Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks

A Thesis
Presented to
The Academic Faculty

by

Guocong Song

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Electrical and Computer Engineering
Georgia Institute of Technology
August 2005

Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks

Approved by:

Professor Ye (Geoffrey) Li, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor James McClellan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Ian F. Akyildiz
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Xingxing Yu
School of Mathematics
Georgia Institute of Technology

Professor John R. Barry
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: April 12, 2005

For Mom and Dad

ACKNOWLEDGEMENTS

My thesis advisor, Professor Ye (Geoffrey) Li, has my sincerest gratitude for having taught me so much about doing research. It was his insight that helped me to initiate this topic. This thesis has significantly benefited from his valuable guidance and constant encouragement. I also greatly appreciate his care and concern for my intellectual and personal growth.

I would like to thank my thesis committee members, Professor Ian F. Akyildiz, Professor John R. Barry, Professor James McClellan, and Professor Xingxing Yu. Their broad perspective and suggestions have helped me in refining this thesis. I appreciate Professor Raghupathy Sivakumar attending my defense and presenting valuable comments.

I want to thank Professor Leonard J. Cimini, Jr. at the University of Delaware and Dr. Haitao Zheng at Microsoft Research for their thoughtful guidance and insightful suggestions.

I would also like to thank Professor Ke Gong at Tsinghua University for shaping my interests in wireless networks.

I am very thankful for my officemates at the Information Transmission and Processing Laboratory, Jingnong Yang, Taewon Hwang, Dr. Hua Zhang, Jianxuan Du, Jet Zhu, Uzoma Anaso Onunkwo, Ghurumuruhan Ganesan, and Wen Jiang. I thank Taewon Hwang for being a great office neighbor for four years, Jianxuan Du for sharing views on almost anything, and Ghurumuruhan Ganesan for intensive technical discussions. I am so proud of being a part of the brilliant and fruitful group.

I also want to thank Yingqun Yu at the University of Minnesota for many inspiring discussions. His intellectual curiosity is always a source of inspiration for me.

Last but not least, I would like to thank my parents for their love, encouragement, and support. This thesis is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Besides avoiding inter-symbol interference and leading to high capacity, wireless *orthogonal frequency division multiplexing* (OFDM) or other multicarrier systems provide fine granularity for resource allocation since they are capable of dynamically assigning subcarriers to multiple users and adaptively allocating transmit power. The current dominate layered networking architecture, in which each layer is designed and operated independently, results in inefficient and inflexible resource use in wireless networks due to the nature of the wireless medium, such as time-varying channel fading, mutual interference, and topology variations. Thus, we need an integrated adaptive design across different layers.

In this thesis, we focus on resource allocation and scheduling in wireless OFDM networks based on joint physical and *medium access control* (MAC) layer optimization. To achieve orders of magnitude gains in system performance, we use two major mechanisms in resource management: exploiting the time variance and frequency selectivity of wireless channels through adaptive modulation, coding, as well as packet scheduling and regulating resource allocation through network economics. With the help of utility functions that capture the satisfaction level of users for a given resource assignment, we establish a utility optimization framework for resource allocation in OFDM networks, in which the network utility at the level of applications is maximized subject to the current channel conditions and the modulation and coding techniques employed in the network. Although the nonlinear and combinatorial nature of the cross-layer optimization challenges algorithm development, we propose novel efficient *dynamic subcarrier assignment* (DSA) and *adaptive power allocation* (APA) algorithms that are proven to achieve the optimal or near-optimal performance with very low complexity. Based on a holistic design principle, we design *max-delay-utility* (MDU) scheduling, which senses both channel and queue information. The MDU scheduling can simultaneously improve the spectral efficiency and provide right incentives to ensure that all applications can receive their different required *quality of service* (QoS). To facilitate

the cross-layer design, we also deeply investigate the mechanisms of channel-aware scheduling, such as efficiency, fairness, and stability. First, using extreme value theory, we analyze the impact of multiuser diversity on throughput and packet delay. Second, we reveal a generic relationship between a specific convex utility function and a type of fairness. Third, with rigorous proofs, we provide a method to design cross-layer scheduling algorithms that allow the queueing stability region at the network layer to approach the ergodic capacity region at the physical layer.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The allocation and management of resources are crucial for wireless networks, in which the scarce wireless spectral resources are shared by multiple users. In the current dominate layered networking architecture, each layer is designed and operated independently to support transparency between layers. Among these layers, the physical layer is in charge of raw-bit transmission, and the *medium access control* (MAC) layer controls multiuser access to the shared resources. However, wireless channels suffer from time-varying multipath fading; moreover, the statistical channel characteristics of different users are different. The suboptimality and inflexibility of this architecture result in inefficient resource use in wireless networks. We need an integrated adaptive design across different layers. Therefore, cross-layered design and optimization across the physical and MAC layers are desired for wireless resource allocation and packet scheduling [3, 62].

For cross-layer optimization, channel-aware scheduling strategies are proposed to adaptively transmit data and dynamically assign wireless resources based on *channel state information* (CSI). The key idea of channel-aware scheduling is to choose a user with good channel conditions to transmit packets [76]. Taking advantage of the independent channel variation across users, channel-aware scheduling can substantially improve the network performance through *multiuser diversity*, whose gain increases with the number of users [33,76]. To guarantee fairness for resource allocation and exploit multiuser diversity, utility-pricing structures in network economics are usually preferred for scheduling design [41].

The growth of Internet data and multimedia applications requires high-speed transmission and efficient resource allocation. To avoid inter-symbol interference, *orthogonal frequency division multiplexing* (OFDM) is desirable for wireless high-speed communications [16]. OFDM-based systems are traditionally used for combating frequency-selective

fading. From a resource allocation point of view, however, multiple channels in an OFDM system naturally have the potential for more efficient MAC since subcarriers can be assigned to different users [15, 79]. Another advantage of OFDM is that adaptive power allocation can be applied for a further improvement.

The basic problem that we need to solve in this thesis is how to effectively allocate resources on the downlink of *Internet protocol* (IP)-based OFDM networks by exploiting knowledge of CSI and the characteristics of traffic to enhance the spectral efficiency and guarantee *quality of service* (QoS).

The objective of this thesis is to establish a theoretical framework and to develop efficient algorithms for resource allocation in wireless multicarrier networks based on cross-layer optimization. This research focuses on both studies on the mechanisms of efficiency, fairness, as well as QoS provisioning and algorithm development for resource allocation in multiuser frequency-selective fading environments.

1.2 Background and Related Work

In this section, we review state-of-the-art techniques for wireless resource allocation, including multiuser diversity, resource allocation in OFDM networks, and network economics.

1.2.1 Multiuser Diversity and Opportunistic Communications

Recently, the principles of multiuser downlink or MAC designs have been changed from the traditional point-to-point view to a multiuser network view. Time-varying fading is a unique characteristic of wireless channels. For a point-to-point link, using adaptive modulation and coding [23, 48], the transmitter can send more data at a higher transmission data rate when the channel quality is good. However, the bandwidth efficiency is still low during deep-fading periods. In [33], the authors have studied the sum capacity of uplink (many-to-one) fading channels in a scenario where the CSI is known for the transmitters and the receiver. There are two important results obtained in [33]. First, the optimal strategy is to choose only one user with the best channel condition. Second, the sum capacity increases with the number of users, which is called *multiuser diversity*. The similar results have been shown in downlink (one-to-many) fading channels in [75].

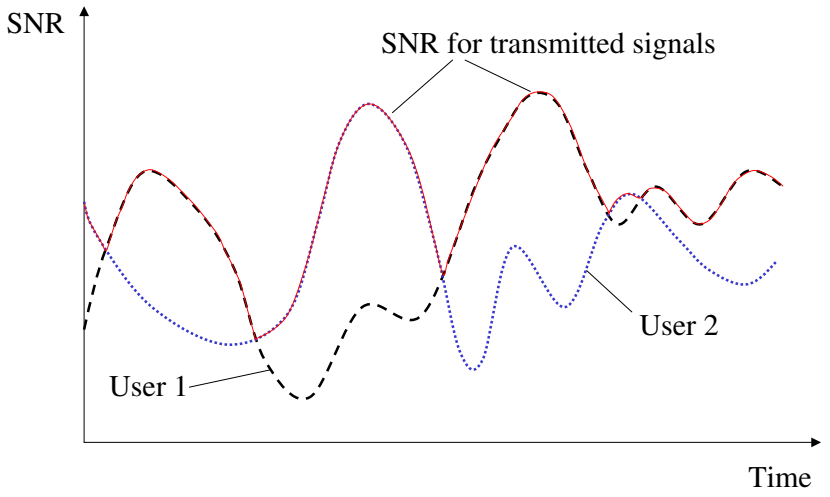


Figure 1.1. Scheduling for the two-user case.

The above results regarding multiuser diversity indicate that the use of simple scheduling techniques and the feedback of CSI can significantly improve spectral efficiency [76]. Actually, multiuser diversity results from the independent channel variation across users. To illustrate multiuser diversity, we consider a two-user case in Figure 1.1, where the user with the best channel condition is scheduled to transmit signals. Therefore, the equivalent *signal-to-noise ratio* (SNR) for transmission is $\max\{\mathrm{SNR}_1(t), \mathrm{SNR}_2(t)\}$. When there are many users served in the system, the packets are with a high probability transmitted at high data rates since different users experience independent fading fluctuations. From a user point of view, packets are transmitted in a stochastic way in the system using channel-aware scheduling, which is also called opportunistic communications [41].

Currently, multiuser diversity has received much attention. Based on its concept, channel-aware dynamic packet scheduling is applied in 1x *evolution* (1xEV) for *code division multiple access 2000* (CDMA2000), also known as IS-856 [73] and *high speed downlink packet access* (HSPDA) for wideband CDMA [2]. Aside from cellular networks, multiuser diversity is also exploited in distributed systems [52, 58].

1.2.2 Resource Allocation for OFDM-Based Networks

OFDM divides an entire channel into many orthogonal narrowband subchannels (subcarriers) to deal with frequency-selective fading and to support a high data rate. Furthermore, in an OFDM-based wireless network, different subcarriers can be allocated to different users to provide a flexible multiuser access scheme [15, 35] and exploit multiuser diversity.

There is plenty of room to exploit the high degree of flexibility of radio resource management in the context of OFDM. Since channel frequency responses are different at different frequencies and for different users, data rate adaptation over each subcarrier, *dynamic subcarrier assignment* (DSA), and *adaptive power allocation* (APA) can significantly improve the performance of OFDM networks. Using data rate adaptation [23, 48], the transmitter can send higher transmission rates over the subcarriers with better conditions so as to improve throughput and simultaneously to ensure an acceptable *bit-error rate* (BER) at each subcarrier. Despite the use of data rate adaptation, deep fading on some subcarriers still leads to low channel capacity.

On the other hand, channel characteristics for different users are almost mutually independent in multiuser environments; the subcarriers experiencing deep fading for one user may not be in a deep fade for other users; therefore, each subcarrier could be in a good condition for some users in a multiuser OFDM wireless network. By dynamically assigning subcarriers, the network can benefit from multiuser diversity. Resource allocation issues and the achievable regions for multiple access and broadcast channels have been investigated in [74] and [36], respectively, which have proved that the largest data rate region is achieved when the same frequency range is shared with overlap by multiple users in broadcast channels. However, when optimal power allocation is used, from [24], there is only a small range of frequency with overlapping power sharing. Thus, optimal power allocation with dynamic subcarrier (non-overlap) assignment can achieve a data transmission rate close to the channel capacity boundary. In [79], the authors have investigated optimal resource allocation in multiuser OFDM systems to minimize the total transmission power while satisfying a minimum rate for each user. The numerical optimization algorithms have been proposed in [83] for characterizing the uplink rate region achievable in OFDM with

inter-symbol interference. Several algorithms have been presented in [32, 55] for subcarrier and power allocation.

1.2.3 Network Economics for Resource Allocation

As seen in previous sections, exploiting multiuser diversity can significantly improve the spectral efficiency. In addition to the spectral efficiency, fairness and QoS are crucial for resource allocation for wireless networks. Usually, it is impossible to achieve the optimality for spectral efficiency, fairness, and QoS simultaneously. For instance, scheduling schemes aiming to maximize the total throughput are unfair to those users far away from a base station or with bad channel conditions. On the other hand, the absolute fairness may lead to low bandwidth efficiency. Therefore, an effective trade-off among efficiency, fairness, and QoS is desired in wireless resource allocation.

The issues on efficient and fair resource allocation have been well studied in economics, where utility functions are used to quantify the benefit of usage of certain resources. Similarly, utility theory can be used in communication networks to evaluate the degree to which a network satisfies service requirements of users' applications, rather than in terms of system-centric quantities like throughput, outage probability, packet drop rate, power, etc. [65]. The basic idea of utility-pricing structures is to map the resource use (bandwidth, power, etc.) or performance criteria (data rate, delay, etc.) into the corresponding utility or price values and optimize the established utility-pricing system.

In wireline networks, utility and pricing mechanisms have been used for flow control [30, 31], congestion control [43], and routing [5]. In wireless networks, the pricing of uplink power control in CDMA has been investigated in [25, 59, 61, 80]. Utility-based power allocation on CDMA downlinks for voice and data applications has been proposed in [40, 69, 84]. To guarantee QoS and exploit multiuser diversity, utility-pricing structures are applied in opportunistic communications [41].

In summary, network economics is becoming more and more important in modern network designs, especially for cross-layer optimization in wireless networks.

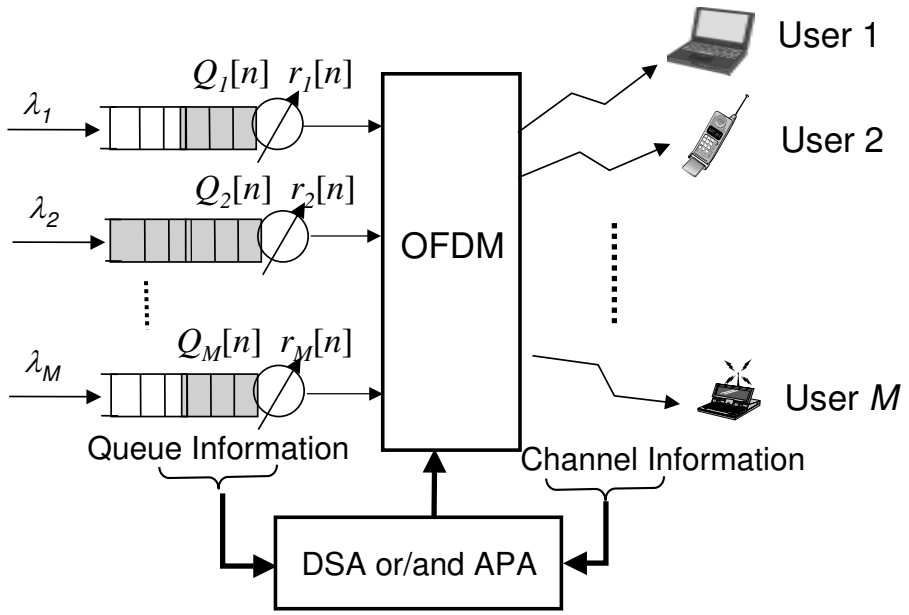


Figure 1.2. Downlink data scheduling over multiple shared channels based on OFDM

1.3 System Model and Problem Description

The architecture of a downlink data scheduler with multiple shared channels for multiple users is shown in Figure 1.2. OFDM provides a physical basis for the multiple shared channels, where the total bandwidth B is divided into K subcarriers (subchannels), and each subcarrier has a bandwidth of $\triangle f = B/K$. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the subcarrier index set. The OFDM signaling is time-slotted, and the length of each time slot is T_s. The base station simultaneously serves M users, each of which has a queue to receive its incoming packets. Let $\mathcal{M} = \{1, 2, \cdots, M\}$ denote the user index set. To achieve high efficiency, both frequency and time multiplexing are allowed in the whole resource. The scheduler makes a subcarrier assignment once every slot based on each user's current channel quality and queue length.

1.3.1 Channel Characteristics in OFDM

The complex baseband representation of the impulse response of a wireless multipath channel for user i can be described by

$$h_i(t, \tau) = \sum_k \gamma_{k,i}(t)\delta(\tau - \tau_{k,i}),$$

where $\tau_{k,i}$ is the delay of the k-th path and $\gamma_{k,i}(t)$ is the corresponding complex amplitude at time t. The $\gamma_{k,i}(t)$'s are assumed to be wide-sense stationary and narrowband stochastic Gaussian processes, which are independent for different paths and users. The frequency response of the above channel impulse response is expressed as

$$H_i(f,t) = \int_{-\infty}^{+\infty} h_i(t,\tau) e^{-j2\pi f\tau}\, d\tau = \sum_k \gamma_{k,i}(t) e^{-j2\pi f \tau_{k,i}}.$$

For OFDM systems with proper cyclic extension and sample timing, the channel frequency response at subcarrier k at time n can be expressed as

$$H_i[k,n] \triangleq H_i(k\triangle f, nT_s).$$

Then, the channel quality of user i is given by

$$\rho_i[k,n] = \frac{|H_i[k,n]|^2}{N_i[k]},$$

where $N_i[k]$ is the noise power of user i at subcarrier k. With a power allocation $p[k,n]$, the SNR at subcarrier k at time n is

$$\gamma_i[k,n] = p[k,n]\rho_i[k,n].$$

There are many ways to obtain the CSI at the base station. In a *frequency division duplex* (FDD) system, using pilot symbols that are inserted in the downlink with a certain time-frequency pattern, the mobile terminals can effectively estimate the channel parameters $H_i[k,n]$'s and $\rho_i[k,n]$'s [37] and feed back them to the base station. In a *time division duplex* (TDD) system, since the symmetry of the channel characteristics for the downlink and uplink, the base station can obtain the CSI by directly measuring the uplink channels.

1.3.2 Rate Adaptation in OFDM

By estimating the CSI via pilot signals and feeding it back to the base station, the achievable data transmission rate per Hz for user i at subcarrier k during time slot n, $c_i[k,n]$, can be known at the base station. Usually, the $c_i[k,n]$'s are determined by the current channel SNR, the required BER, and the modulation and coding techniques that are used in the system.

If continuous rate adaptation is used, the achievable transmission rate per Hz at subcarrier k for user i can be written as a function of the current SNR, $\gamma_i[k,n]$, [53]

$$c_i[k,n] = \log_2(1 + \beta\gamma_i[k,n]), \tag{1.1}$$

where β is a constant related to a targeted BER by

$$\beta = \frac{-1.5}{\ln(5 \cdot \text{BER})}.$$

Generally, $c_i[k,n]$ can be expressed as

$$c_i[k,n] = f\left(\log_2(1 + \beta\gamma_i[k,n])\right), \tag{1.2}$$

where $f(\cdot)$ depends on the used rate adaptation scheme. For instance, if variable *M-ray quadrature amplitude modulation* (MQAM) with modulation levels {0, 2, 4, 6,...} is employed,

$$f(x) = 2\lfloor\frac{1}{2}x\rfloor,$$

where $\lfloor x \rfloor$ represents the largest integer that is less than x.

1.3.3 Dynamic Subcarrier Assignment and Adaptive Power Allocation

Each subcarrier in the adaptive OFDM can be dynamically assigned to any user. Let $D_i^{(n)}$ denote the set of subcarrier indices assigned to user i at time n. In the OFDM system, each subcarrier cannot be shared by multiple users, which is mathematically expressed as

$$D_i^{(n)} \bigcap D_j^{(n)} = \varnothing, \quad \forall\ i \neq j,$$
$$\bigcup_{i \in \mathcal{M}} D_i^{(n)} \subseteq \mathcal{K}.$$

With a subcarrier assignment, the data transmission rate of user i at time slot n, $r_i[n]$, is given by

$$r_i[n] = \sum_{k \in D_i^{(n)}} c_i[k,n] \triangle f.$$

Let $\mathbf{p}[n]$ be the transmit power vector defined as $[p[1,n], p[2,n], \ldots, p[K,n]]^T$, which $p[k,n]$ is the transmit power at subcarrier k at time n. If the adaptive power allocation is used, the transmit powers can be adjusted but constrained by

$$\sum_{k=1}^{K} p[k,n] \leq \bar{P},$$

where $\bar{P}$ is the total power constraint.

1.3.4 Queue Structure

Each connection is assumed to have a queue with infinite capacity at the base station. Let $Q_i[n]$ be the amount of bits in the queue of user i at time nT_s. During time slot n, the base station serves the queue of user i at rate $r_i[n]$. Then, the queue length evolution equation is given by

$$Q_i[n+1] = Q_i[n] - \min\left(Q_i[n], r_i[n]T_s\right) + a_i[n] \tag{1.3}$$

where $a_i[n]$ is the amount of arrival bits during time slot n.

1.3.5 Problem Description

The major problem is how to effectively assign subcarriers and allocate power on the downlink of OFDM-based networks by exploiting knowledge of the wireless channel conditions and the characteristics of traffic to improve the spectral efficiency and guarantee diverse QoS.

1.4 Our Approach

In this thesis, we primarily take an analytical approach and use simulation to support the theoretical results and demonstrate the performance of ours schemes in more realistic environments. In the joint physical and MAC layer optimization framework, we use two major mechanisms in resource management: exploiting the time variance and the frequency selectivity of wireless channels in network protocols through adaptive modulation, coding, as well as packet scheduling and regulating resource allocation through network economics. Besides leading to high capacity, OFDM provides fine granularity for resource allocation since different subcarriers can be assigned to different users. With the help of utility functions that capture the satisfaction level of users for a given resource assignment, we establish a utility optimization framework for resource allocation in OFDM networks, in which the network utility at the level of applications is maximized subject to the channel conditions and the modulation and coding techniques employed in the network.

Figure 1.3 illustrates the structure of the thesis. In the cross-layer optimization with utility functions with respect to instantaneous data rates, we develop novel efficient DSA and APA algorithms, which provide the algorithm implementation for channel-aware scheduling and joint channel- and queue-aware scheduling. Using utility functions with respect to average data rates, we can design channel-aware scheduling desirable for best-effort traffic. We reveal a generic relationship between a specific convex utility function and a type of fairness. Based on a holistic design principle, we develop a joint channel- and queue-aware scheduling scheme that maximizes the total utility with respect to average delays. The stability issue of the queueing system is comprehensively investigated because of the importance to delay-sensitive traffic. The utility-based architecture is finally proven to have the ability of QoS differentiation for heterogeneous traffic. Moreover, in the case when the utility function is just the throughput, we provide a concise asymptotic analysis for throughput and packet delay to reveal the impact of multiuser diversity.

From a traditional point of view, cross-layer design would usually seem complicated and intractable. Therefore, we pursue consistency in methodology and simplicity in results in this thesis. Asymptotic approach is extensively used in analysis since it leads to elegant results. Those results can be not only helpful in obtaining insights but also fully applied to the system design. For instance, the study on the optimization properties of the "extreme" OFDM in which the number of subcarrier is infinite directly guides the algorithm development for practical systems. The study on the stability properties of joint channel- and queue-aware scheduling plays a crucial role in designing scheduling for heterogeneous traffic with diverse QoS requirements. The asymptotic throughput analysis of channel-aware scheduling is very accurate for typical environments and can deal with a general fading distribution. In addition, the convexity of the ergodic capacity region at the physical layer is fully exploited throughout the thesis, which makes most results concerning such complicated problems as fairness and stability elegant.

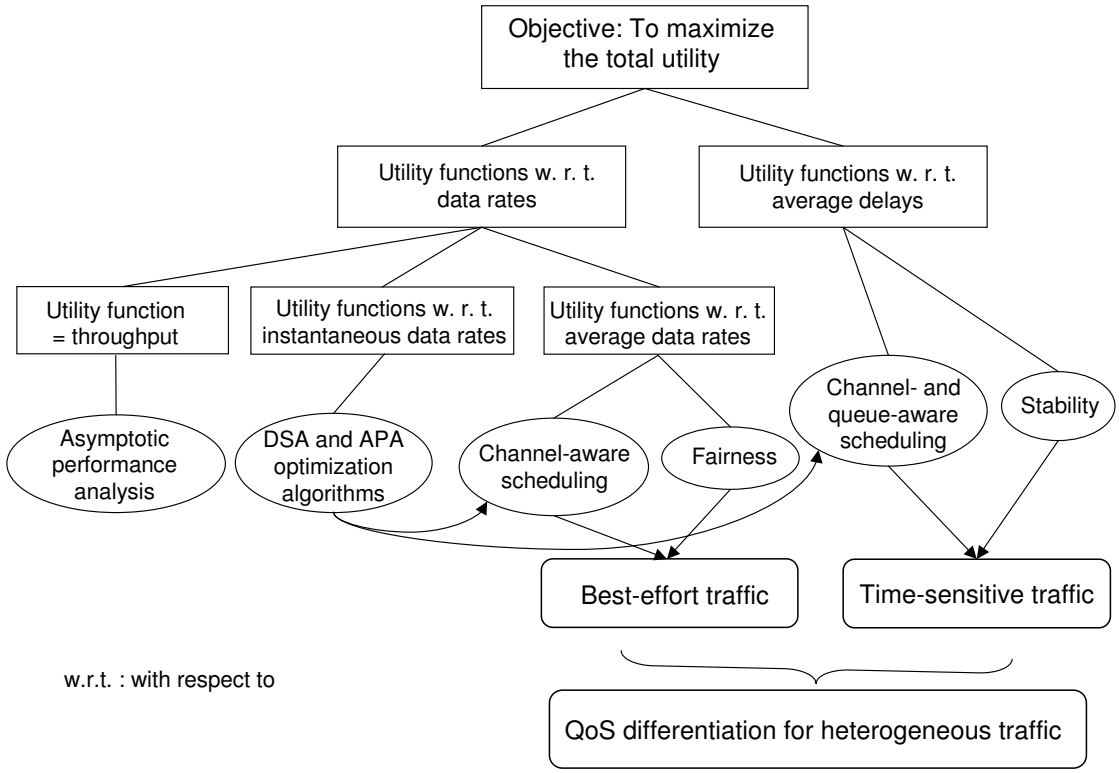


Figure 1.3. Structure of the thesis research

1.5 Thesis Outline

The thesis is organized as follows. In Chapter 2, using rate-based utility functions, we formulate the cross-layer optimization problem as one that maximizes the total utility of all active users subject to certain conditions, which are determined by adaptive resource allocation schemes. We present the necessary and sufficient conditions for the utility-based optimal subcarrier assignment and power allocation for the asymptotic case in which the number of subcarrier is infinite. Taking various conditions into account, we develop a variety of efficient algorithms, including sorting-search dynamic subcarrier assignment, greedy bit loading and power allocation, as well as objective aggregation algorithms. We also modify those algorithms for a certain type of non-concave utility function. In addition, with utility functions with respect to average data rates, time diversity can be exploited to further improve performance.

In Chapter 3, packet scheduling in a shared multicarrier downlink is investigated based

on cross-layer design and optimization. We first develop *max-delay-utility* (MDU) scheduling, a joint channel- and queue-aware scheduling scheme that maximizes the total utility with respect to average delays, to exploit multiuser diversity and guarantee QoS. The stability property of a scheduling policy is characterized by the stability region, which is the largest region on the arrival rates for which the queueing system can be stabilized by the scheduling policy. It is shown that under very loose conditions, the MDU scheduling has the maximum stability region. In environments with insufficient scattering or strong light-of-sight components, delay transmit diversity can increase the fluctuation in the frequency domain, thereby improving the performance.

In Chapter 4, we use the MDU scheduling to effectively provide QoS differentiation for heterogenous traffic. The mechanisms of the MDU scheduling at resource allocation and stability aspects are discussed in the scenario in which multiple types of traffic are served. A comprehensive simulation that takes into account packet-switched voice, streaming, and best-effort applications demonstrates the advantages of the MDU scheduling for integrated services with diverse QoS.

In Chapter 5, we provide an asymptotic performance analysis of channel-aware packet scheduling based on extreme value theory. We first address the average throughput of systems with a homogeneous average SNR and obtain its asymptotic expression. Compared to the exact throughput expression, the asymptotic one, which is applicable to a broader range of fading channels, is more concise and easier to get insights. For a system with heterogeneous SNRs, normalized-SNR-based scheduling need to be used for fairness. We investigate the asymptotic average throughput of the normalized-SNR-based scheduling and prove that the average throughput in this case is less than that in the homogeneous case with a power constraint. Furthermore, we propose an asymptotic delay analysis for both single-carrier and multicarrier systems based on extreme value theory and queueing theory. The asymptotic analysis for mean packet delays demonstrates that the multiuser diversity gain in multicarrier networks is not limited by slow fading as in single-carrier networks.

CHAPTER 2

CROSS-LAYER RESOURCE ALLOCATION AND SCHEDULING USING RATE-BASED UTILITY FUNCTIONS

In this chapter, we do not consider the burstness of arrival streams and investigate resource allocation and scheduling for best-effort traffic. Therefore, rate-based utility functions are used to perform cross-layer optimization and balance efficiency and fairness in this chapter. In Section 2.1, we the general properties of rate-based utility functions. In Sections 2.2 and 2.3, we investigate the optimization problems at an instantaneous time, which are formulated based on utility functions with respect to instantaneous data rates. In Section 2.2, we focus on the OFDM network that contains infinite number of subcarriers and employs continuous rate adaptation. In Section 2.2, we develop efficient resource allocation algorithms for the cross-layer optimization in various system configurations. In Section 2.4, we propose channel-aware scheduling based on utility functions with respect to average data rates. The algorithms developed in Section 2.2 can be directly used for the channel-aware scheduling. In Section 2.5, we discuss the efficiency and fairness issues. The relationship between a utility function and a certain type of fairness is revealed. In Section 2.6, we demonstrate the performance improvement of the cross-layer optimization through numerical results.

2.1 Rate-Based Utility Functions

Utility functions are used for the cross-layer optimization and balancing the efficiency and fairness. A utility function maps the network resources that a user utilizes into a real number. In almost all wireless applications, a reliable data transmission rate is the most important factor to determine the satisfaction of users. Thus, the utility function $U(r)$ should be a nondecreasing function of the data rate r. In particular, when $U(r) = r$, the

utility is just the throughput, which is the objective of most traditional network optimizations. Therefore, our work can be regarded as a general extension of traditional network optimizations.

Utility functions serve as an optimization objective for the adaptive physical and MAC layer techniques. Consequently, they can be used to optimize radio resource allocation for different applications and to build a bridge among the physical, MAC, and higher layers.

When a utility function is used to capture the user's feeling, such as the level of satisfaction for assigned certain resources, it cannot be obtained only through theoretical derivation. In this case, it can be estimated from subjective surveys. For best-effort traffic [29], a utility function can be described by

$$U(r) = 0.16 + 0.8\ln(r - 0.3), \tag{2.1}$$

where r is in unit of kbps. To prevent assigning too much resource to the user with good channel conditions, the slope of the utility curves decreases with an increase in the data rate. We will discuss more on the issue of fairness and efficiency in Section 2.5.

2.2 Theoretical Framework

To obtain the performance bound of the cross-layer optimization, we assume in this section that there is an infinite number of orthogonal subcarriers in all frequency resources, or the bandwidth of each orthogonal subcarrier is infinitesimal, which can be regarded as an extreme situation of OFDM. In a practical OFDM system, the minimum granularity of resource allocation is one subcarrier. The OFDM system in which $\triangle f \rightarrow 0$ provides an infinitesimal granularity of resource allocation, thereby presenting the performance upper bound.

2.2.1 Problem Formulation

Since we consider the "extreme" OFDM system, some parts of the system model in Section 1.3 should be modified slightly. Thus, we will briefly describe the modifications in the system model.

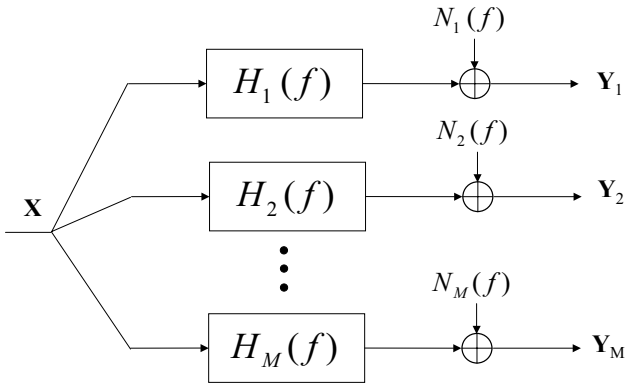


Figure 2.1. Channel model

Because we investigate the cross-layer optimization in terms of the instantaneous data rates, we ignore time parameter t in all formulas in this section.

The M-user frequency-selective broadcast fading channel is shown in Figure 2.1. The channel frequency response corresponding to user i is denoted by $H_i(f)$. The quality of each user's channel can be indicated by the SNR function, $\rho_i(f)$, when the transmission power density $p(f) = 1$, which is defined as

$$\rho_i(f) = \frac{\mid H_i(f) \mid^2}{N_i(f)}.$$

where $N_i(f)$ is the noise power density function of user i.

Let $c_i(f)$ denote the achievable throughput of user i at frequency f for a given BER and a transmission power density $p(f)$. When continuous rate adaptation is used, $c_i(f)$ can be expressed as [53]

$$\begin{aligned} c_i(f) &= \log_2(1 + \frac{\beta p(f) \mid H_i(f) \mid^2}{N_i(f)}) \quad \text{(bits/sec/Hz)} \\ &= \log_2(1 + \beta p(f)\rho_i(f)). \end{aligned} \tag{2.2}$$

In the scenario, the D_i's become the frequency sets assigned to different users, which are constrained by

$$D_i \bigcap D_j = \varnothing, \quad i \neq j, \tag{2.3}$$

$$\bigcup_{i=1}^{M} D_i \subseteq [0, B]. \tag{2.4}$$

Besides dynamically assigning the frequency sets, the transmission power density at different frequencies can also be adjusted to improve the network performance with a total transmission power constraint by

$$\frac{1}{B}\int_0^B p(f)\,df \le 1. \tag{2.5}$$

The achievable transmission efficiency of user i in the continuous-frequency case is given by

$$c_i(f) = \log_2[1 + \beta p(f)\rho_i(f)],$$

and the transmission throughput of user i can be expressed as

$$r_i = \int_{D_i} c_i(f)\,df. \tag{2.6}$$

Let the utility function of user i be $U_i(\cdot)$. If user i has a data rate r_i, the user's utility is $U_i(r_i)$. The utility-based cross-layer optimization is to assign wireless resources (including frequency band and power density) to maximize the average utility of the network, which can be expressed as

$$\frac{1}{M}\sum_{i=1}^{M} U_i(r_i). \tag{2.7}$$

In the next several sections, we will discuss *dynamic subcarrier assignment* (DSA), *adaptive power allocation* (APA), and joint DSA and APA, respectively.

2.2.2 Dynamic Subcarrier Assignment

In this section, we investigate DSA to improve the performance of an OFDM-based network when the transmission power is uniformly distributed over the entire available frequency band, that is, $p(f) = 1$, then the achievable throughput at frequency f, $c_i(f)$, can be expressed as

$$c_i(f) = \log_2(1 + \beta\rho_i(f)).$$

Thus, the DSA problem is to maximize

$$\frac{1}{M}\sum_{i=1}^{M} U_i(r_i) = \frac{1}{M}\sum_{i=1}^{M} U_i\left(\int_{D_i} c_i(f)\,df\right), \tag{2.8}$$

subject to

$$\bigcup_{i=1}^{M} D_i \subseteq [0, B], \tag{2.9}$$

$$D_i \bigcap D_j = \varnothing, \quad i \neq j \text{ and } i,\ j = 1,\ 2, \cdots, M. \tag{2.10}$$

We first present the results for a network with two users and then extend to general networks.

2.2.2.1 Network with Two Users

Assume a network with only 2 users sharing the bandwidth $[0,\ B]$. Define

$$\bar{D}_1(\alpha) = \{f \in [0, B] : \frac{c_2(f)}{c_1(f)} = \frac{\log_2(1 + \beta\rho_2(f))}{\log_2(1 + \beta\rho_1(f))} \leq \alpha\}, \tag{2.11}$$

and

$$D_1(\alpha) = \{f \in [0, B] : \frac{c_2(f)}{c_1(f)} = \frac{\log_2(1 + \beta\rho_2(f))}{\log_2(1 + \beta\rho_1(f))} < \alpha\}. \tag{2.12}$$

Similarly, we can define $\bar{D}_2(\alpha)$ and $D_2(\alpha)$ as the regions where $\frac{c_2(f)}{c_1(f)} >= \alpha$ and $\frac{c_2(f)}{c_1(f)} > \alpha$, respectively. It can be easily seen that

$$\bar{D}_2(\alpha) \cup D_1(\alpha) = \bar{D}_1(\alpha) \cup D_2(\alpha) = [0, B],$$

and

$$\bar{D}_2(\alpha) \cap D_1(\alpha) = \bar{D}_1(\alpha) \cap D_2(\alpha) = \varnothing.$$

The following theorem is proved in Appendix A and it determines the optimal subcarrier assignment for the cross-layer optimization.

Theorem 2.1 *For a network with two users, if the subcarrier assignment, $\{D_1^*, D_2^*\}$, is optimal, then D_1^* and D_2^* satisfy*

$$D_1(\alpha^*) \subseteq D_1^* \subseteq \bar{D}_1(\alpha^*), \quad D_2^* = [0,\ B] - D_1^*, \quad \alpha^* = \frac{U_1'(r_1^*)}{U_2'(r_2^*)},$$

and

$$r_i^* = \int_{D_i^*} c_i(f)df = \int_{D_i^*} \log_2(1 + \beta\rho_i(f))df, \text{ for } i = 1,\ 2,$$

where

$$U_i'(r) = \frac{dU_i(r)}{dr}.$$

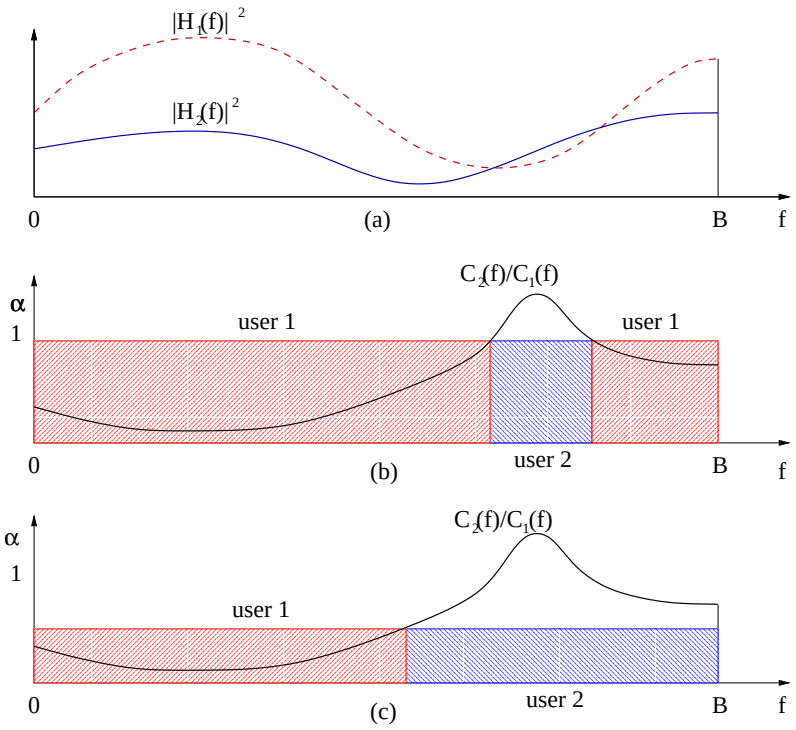


Figure 2.2. Optimal subcarrier assignment for a two-user network

(a) Frequency responses for two users; (b) Subcarrier assignment resulting from throughput-based optimization; (c) Subcarrier assignment resulting from utility-based optimization

Figure 2.2 demonstrates the difference between the utility-based optimization and the traditional throughput-based optimization. For the traditional optimization, $U_i(r) = r$; therefore, the threshold, $\alpha^* = \frac{U_1'(r_1)}{U_2'(r_2)}$ is always 1. Consequently, a subcarrier or frequency is allocated to the user with the larger channel gain, as in Figure 2.2 (b). To balance the efficiency and fairness, an increasing utility curve with a decreasing slope is usually used. In this case, the threshold α^* depends on how much resource each user has already occupied. Since the channel corresponding to user 2 is not as good as that of user 1 in Figure 2.2 (a), user 2 gets more frequency resource in the utility-based optimization than in the throughput-based optimization, as in Figure 2.2 (c).

It should be noted that the optimal subcarrier assignment is not unique as we can see in a network with flat fading channels. However, α^*, r_1^*, and r_2^* are unique.

2.2.2.2 Network with Multiple Users

The results for a two-user network can be extended to the general case of more than two users, which is summarized in the following theorem.

Theorem 2.2 *For a network with M users, if the subcarrier assignment, D_i^*'s for $i = 1,\ 2, \cdots, M$, maximizes the average utility, then for any $f \in D_i^*$, we have*

$$U_j^{'}(r_j^*)c_j(f) \leq U_i^{'}(r_i^*)c_i(f), \quad \textit{for any } j \neq i, \tag{2.13}$$

and

$$r_i^* = \int_{D_i^*} c_i(f)df.$$

The proof of the above theorem is very similar to that of Theorem 2.1 and is omitted here.

2.2.3 Adaptive Power Allocation

In the previous section, in which the power allocation is assumed to be fixed, we discussed using DSA to maximize the network performance. In this section, we first investigate APA with fixed subcarrier assignment and then study joint DSA and APA. Since the achievable throughput is a function of the power allocation, it becomes

$$c_i(f) = \log_2(1 + \beta p(f)\rho_i(f)).$$

2.2.3.1 Adaptive Power Allocation with Fixed Subcarrier Assignment

When a subcarrier assignment is fixed, the APA optimization can be formulated as follows: given a fixed subcarrier assignment, D_i's for $i = 1,\ 2, \cdots, M$, assign the power density, $p(f)$, to maximize

$$\frac{1}{M}\sum_{i=1}^{M} U_i(r_i) = \frac{1}{M}\sum_{i=1}^{M} U_i\left(\int_{D_i} \log_2[1 + \beta p(f)\rho_i(f)]\, df\right), \tag{2.14}$$

subject to

$$\frac{1}{B}\int_0^B p(f)\, df \leq 1, \quad \text{and} \quad p(f) \geq 0. \tag{2.15}$$

To achieve its optimality, a utility-based multi-level water filling is needed, which is stated in the following theorem.

Theorem 2.3 *For a given fixed subcarrier assignment, D_i's for all i, the optimal power allocation, $p^*(f)$, satisfies*

$$p^*(f) = \left[\frac{U_i^{'}(r_i^*)}{\lambda} - \frac{1}{\beta\rho_i(f)}\right]^+ \qquad \lambda > 0, \quad f \in D_i \tag{2.16}$$

where λ is a constant for the normalization of the optimal power density,

$$[x]^+ = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases},$$

and λ as well as the r_i^*'s satisfy

$$\frac{1}{B}\int_0^B p^*(f)\,df = 1,$$
$$\text{and} \qquad r_i^* = \int_{D_i} \log_2[1 + \beta p^*(f)\rho_i(f)]\,df,$$

where the r_i^*'s and $p^*(f)$ are the optimal values of the rates and the power density, respectively.

It should be indicated that Theorem 2.3 only gives a necessary condition for the globally optimal power allocation. The proof of the above theorem is similar to the water-filling theorem [64], which is summarized in Appendix B.

Similar to the classical water filling [64], the optimal power allocation cannot be directly calculated from (2.16), and iterative algorithms are needed to obtain the optimal one satisfying the power constraint.

There are two major differences between the classical water-filling and the one in Theorem 2.3. First, the *water level* for each user is proportional to its current marginal utility value, $U_i^{'}(r_i)$. In other words, the power allocation is also related to the utility functions. Since the data rates of users are unlikely equal, it is from (2.16) that the water levels, $\frac{U_i^{'}(r_i^*)}{\lambda}$'s, are different for different users. Second, the power constraint is the total transmission power rather than the power of an individual user. As shown in Figure 2.3, the utility-based multi-level water-filling (2.16) can be regarded as an extension of the fixed-priority multi-level water-filling in [27].

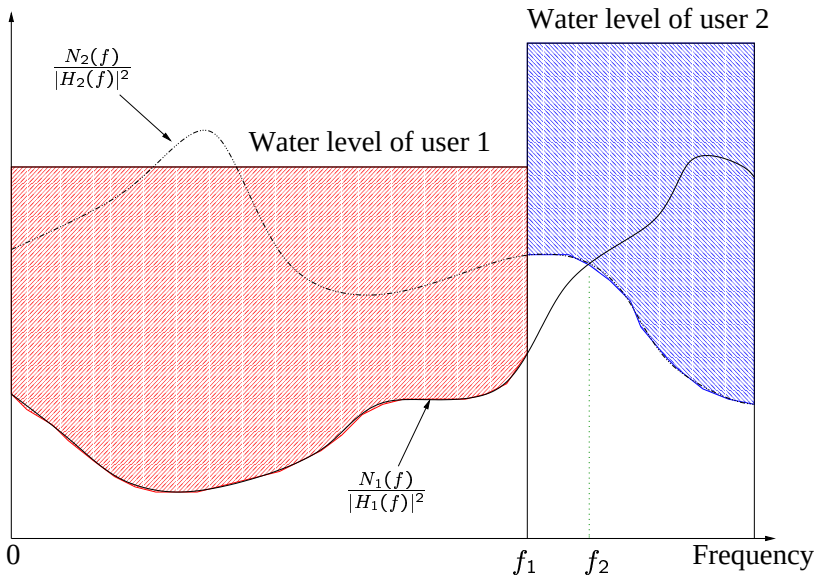


Figure 2.3. Multi-level water-filling for adaptive power allocation in a two-user network.

2.2.3.2 Joint Dynamic Subcarrier Assignment and Adaptive Power Allocation

The DSA and APA can be used simultaneously for the cross-layer optimization. The joint DSA and APA optimization can be formulated as follows: adjust the D_i's and $p(f)$ to maximize

$$\frac{1}{M}\sum_{i=1}^{M} U_i(r_i) = \frac{1}{M}\sum_{i=1}^{M} U_i\left(\int_{D_i} \log_2[1+\beta p(f)\rho_i(f)]\,df\right), \tag{2.17}$$

subject to

$$\bigcup_{i=1}^{M} D_i \subseteq [0, B], \tag{2.18}$$

$$D_i \bigcap D_j = \varnothing, \quad i \neq j \text{ and } i,\ j = 1,\ 2, \cdots, M, \tag{2.19}$$

and

$$\frac{1}{B}\int_0^B p(f)\,df \leq 1 \text{ and } p(f) \geq 0. \tag{2.20}$$

Obviously, there are two necessary conditions for the global optimum for the joint DSA and APA:

1. Fixing the optimal subcarrier assignment, any change of the power allocation does not increase the total utility.

2. Fixing the optimal power allocation, any change of the subcarrier assignment does not increase the total utility.

Therefore, an optimal frequency assignment, D_i^*'s for all i, and power allocation $p^*(f)$ must satisfy the conditions in both Theorems 2.2 and 2.3. Consequently, we have the following theorem.

Theorem 2.4 *Let the D_i^*'s for $i = 1, 2, \cdots, M$ and $p^*(f)$ be the optimal subcarrier assignment and power allocation, respectively. Then they satisfy the following conditions:*

$$\begin{cases} U_j'(r_j^*)\log_2(1+\beta p^*(f)\rho_j(f)) \le U_i'(r_i^*)\log_2(1+\beta p^*(f)\rho_i(f)) & f \in D_i^*, \\ p^*(f) = \left[\dfrac{U_i'(r_i^*)}{\lambda} - \dfrac{1}{\beta\rho_i(f)}\right]^+ & \lambda > 0 \quad f \in D_i^*, \end{cases} \tag{2.21}$$

where the r_i^*'s and λ are constrained by

$$\frac{1}{B}\int_0^B p^*(f)\,df = 1,$$
$$\text{and} \qquad r_i^* = \int_{D_i^*} \log_2(1+\beta p^*(f)\rho_i(f))df.$$

When the utility function is just the throughput, $U_i(r_i) = r_i$, the optimal subcarrier assignment is independent of the optimal power allocation. In this case, the optimal subcarrier assignment and power allocation have the following closed forms:

$$\begin{cases} D_i^* = \{f \in [0:B] : \rho_i(f) = \max_m \rho_m(f)\} \\ p^*(f) = \left[\dfrac{1}{\lambda} - \dfrac{1}{\beta \max_m \rho_m(f)}\right]^+ \\ \dfrac{1}{B}\displaystyle\int_0^B p^*(f)\,df = 1, \end{cases}$$

which is identical to the result in [75]. This illustrates that *frequency division multiple access* (FDMA)-type systems can achieve Shannon capacity when they are optimized for the sum of throughputs.

2.2.4 Properties of Cross-Layer Optimization

In this section, we will prove the convexity of the achievable data rate region and show that, if the utility function is concave, then a local maximum is also a global maximum. Therefore, the necessary conditions in Theorems 2.2, 2.3, and 2.4 are also sufficient ones.

2.2.4.1 Convexity of Instantaneous Data Rate Region

A data rate vector $\mathbf{r}$ is defined as

$$\mathbf{r} = (r_1, r_2, \ldots, r_M)^T \in \mathbb{R}_+^M,$$

where M is the number of users. The *instantaneous data rate region*, $\mathcal{C}_\pi$, is a set that consists of the total achievable data rate vectors under the constraint of a resource allocation policy π, such as DSA, APA, or joint DSA and APA. The instantaneous data rate region is obviously determined by the channel conditions and the resource allocation constraints. It is intuitive that more adaptive resource allocation techniques will result in a larger feasible region.

The objective function is

$$U(\mathbf{r}) = \frac{1}{M} \sum_{i=1}^{M} U_i(r_i).$$

Thus, the optimization problem can be regarded as

$$\max_{\mathbf{r} \in \mathcal{C}_\pi} U(\mathbf{r})$$

Therefore, if $\mathcal{C}_\pi$ is convex, the optimization problem will become tractable. The convexity of the instantaneous data rate region with frequency assignment and power allocation can be described by the following theorem, which is proved in Appendix C.

Theorem 2.5 *For an OFDM-based network with infinitesimal subcarrier space and with DSA, APA, or joint DSA and APA, the achievable data rate region is convex.*

With the above theorem, we can obtain the following property of the cross-layer optimization.

Lemma 2.1 *Let the boundary of the instantaneous data rate region be a subset of the data rate region with the following property: no component of any data rate vector can be increased while the other data rate components remain fixed. The data rate with respect to the maximum of the average utility must be on the boundary of the data rate region if each utility function is strictly increasing.*

Proof: Suppose that the maximum can be achieved by a data rate vector $\mathbf{r}$, which is not on the boundary of the data rate region. There must exist a vector $\mathbf{r}^*$ such that $\mathbf{r} \leqslant \mathbf{r}^*$ [1] with $r_i < r_i^*$ for some i, then $U(\mathbf{r}) < U(\mathbf{r}^*)$. The contradiction shows Lemma 2.1. □

Lemma 2.1 implies that using a strictly increasing utility function intends to assign all resources including all power and bandwidth to users.

2.2.4.2 Global Optimum

For general differentiable utility functions, the conditions (2.13), (2.16), and (2.21) are sufficient and necessary for *locally* optimal solutions of respective optimization problems; hence they are only necessary for the *global* optimality. With concave utility functions, however, the global optimality of the cross-layer optimization can be described by the following theorem.

Theorem 2.6 *If all $U_i(r_i)$'s are concave functions, then a local maximum of $U(\mathbf{r})$ is also a global maximum, and the conditions (2.13), (2.16) and (2.21) are not only necessary but also sufficient, respectively.*

Proof: The proof simply uses the following two consequences in convex analysis [56].

1. If all $U_i(r_i)$ are concave functions, then the objective function $U(\mathbf{r}) = \frac{1}{M}\sum_{i=1}^{M} U_i(r_i)$ is also a concave function.

2. If $\mathcal{C}_\pi \in \mathbb{R}^n$ is a convex set and $U : \mathcal{C}_\pi \mapsto \mathbb{R}$ is a concave function, then a local maximum of U is also a global maximum.

□

[1] $\mathbf{r} \leqslant \mathbf{r}^*$ means $r_i \leq r_i^*$, for all i.

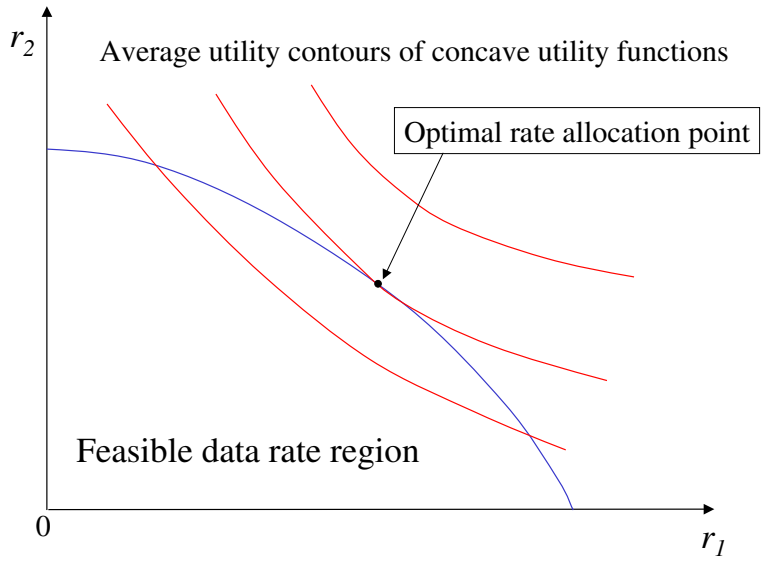


Figure 2.4. Feasible data rate region and optimal rate allocation

The sufficiency of the conditions (2.13), (2.16), and (2.21) for a global optimum is indispensable for algorithm design. If, in addition, the $U_i(r_i)$'s are all strictly concave, there is a unique global maximum solution to the optimization problems. Note that the unique global maximum implies that there is only one optimal data rate vector. However, there may be different frequency and power allocation schemes corresponding to the optimal data rate vector as we can see from a network with flat fading channels for all users.

The relation between the feasible data rate region and concave utility functions is shown in Figure 2.4. Heuristically, Lemma 2.1 shows that the rate vector corresponding to the maximum is located on the boundary of the achievable rate region. Therefore, the optimal rate vector should be a point of tangency between the region boundary and an average utility contour.

2.3 Algorithm Development

We have established a theoretical framework for cross-layer optimization in OFDM wireless networks in Section 2.2. In this section, we focus on effective and practical algorithms for efficient and fair resource allocation in OFDM wireless networks. In practical OFDM wireless networks, the number of subcarriers is finite; therefore, the optimization problem turns from continuous to discrete. This discrete optimization, together with nonlinear utility

functions, challenges algorithm design.

The system model is presented in Section 1.3, but the time parameter n is omitted in this section. Given a power vector $\mathbf{p}$, the achievable transmission efficiency at subcarrier k is denoted as $c_i^{\mathbf{P}}[k]$. To make optimization problems more tractable, we will further assume that the utility curve is continuously differentiable. We take various conditions into account and develop a variety of efficient algorithms, including sorting-search dynamic subcarrier assignment, greedy bit loading and power allocation, as well as objective aggregation algorithms. Furthermore, we will also extend our discussion to a special type of non-concave utility functions in Section 2.4.

The use of utility functions with respect to average data rates can further improve performance by exploiting time diversity. A low-pass time filter can be easily incorporated into all of the algorithms.

2.3.1 Dynamic Subcarrier Assignment Algorithms

In this section, we develop DSA algorithms by assuming a fixed power allocation. When only DSA is used, the problem can be mathematically formulated as follows: given a fixed power allocation, $\mathbf{p}$,

$$\max_{D_i, i\in\mathcal{M}} \quad \sum_{i\in\mathcal{M}} U_i(r_i) \tag{2.22}$$

$$\text{subject to} \quad \bigcup_{i\in\mathcal{M}} D_i \subseteq \mathcal{K}, \tag{2.23}$$

$$D_i \bigcap D_j = \varnothing, \quad i \neq j \quad \forall i, j \in \mathcal{M}, \tag{2.24}$$

Unlike the scenario of infinite subcarriers, which is analyzed in Section 2.2, the instantaneous data rate region determined by (2.23) and (2.24) is not convex anymore. Therefore, it is necessary to investigate the corresponding optimality conditions.

2.3.1.1 Optimality Conditions

In order to study the optimality, we reformulate the above discrete DSA problem as a nonlinear integer (0-1) programming one. Let x_{ik} indicate whether subcarrier k is assigned to

user i or not, that is,

$$x_{ik} = \begin{cases} 1, & \text{if subcarrier } k \text{ is assigned to user } i, \\ 0, & \text{otherwise.} \end{cases}$$

Then the equivalent non-linear integer (0-1) programming problem can be described as follows.

$$\begin{aligned} \max_{\mathbf{x}} \quad & \sum_{i \in \mathcal{M}} U_i \Big(\triangle f \sum_{k \in \mathcal{K}} c_i^{\mathbf{P}}[k] x_{ik} \Big) \\ \text{subject to} \quad & \sum_{i \in \mathcal{M}} x_{ik} = 1, \ k \in \mathcal{K}, \text{ and} \\ & x_{ik} \in \{0, 1\}, \quad i \in \mathcal{M}, \ k \in \mathcal{K}, \end{aligned}$$

where $\mathbf{x} = [x_{11}, \cdots, x_{1K}, x_{21}, \cdots, x_{2K}, \cdots, x_{M1}, \cdots, x_{MK}]^T$. Thus, there is a one-to-one correspondence between $\mathbf{x}$ and the D_i's.

Let

$$U(\mathbf{x}) = \sum_{i \in \mathcal{M}} U_i \Big(\triangle f \sum_{k \in \mathcal{K}} c_i^{\mathbf{P}}[k] x_{ik} \Big),$$

and $\mathcal{B}$ be the feasible region of $\mathbf{x}$'s. If the utility functions $U_i(r)$'s are concave and differentiable, from the property of the subgradient of concave functions [56], $\forall \mathbf{x} \in \mathcal{B}$,

$$U(\mathbf{x}) - U(\mathbf{y}) \geq \nabla_{\mathbf{x}} U(\mathbf{x})^T (\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{B}. \tag{2.25}$$

where the gradient of $U(\mathbf{x})$ is defined as,

$$\nabla_{\mathbf{x}} U(\mathbf{x}) = \begin{bmatrix} U_1'(r_1) c_1^{\mathbf{P}}[1] \triangle f \\ \vdots \\ U_1'(r_1) c_1^{\mathbf{P}}[K] \triangle f \\ \vdots \\ U_M'(r_M) c_M^{\mathbf{P}}[1] \triangle f \\ \vdots \\ U_M'(r_M) c_M^{\mathbf{P}}[K] \triangle f \end{bmatrix},$$

with

$$U_i'(r) = \frac{dU_i(r)}{dr}.$$

It can be directly derived from (2.25) that if $\mathbf{x}^*$ satisfies

$$\nabla_{\mathbf{x}} U(\mathbf{x}^*)^T(\mathbf{x}^* - \mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \mathcal{B}, \tag{2.26}$$

then

$$U(\mathbf{x}^*) - U(\mathbf{x}) \geq 0, \ \forall \mathbf{x} \in \mathcal{B},$$

which means that $\mathbf{x}^*$ is globally optimal. The condition (2.26) is equivalent to the following expression. $\forall k \in \mathcal{K}$, letting i be the user index with respect to k such that $x_{ik}^* = 1$,

$$U_i^{'}(r_i^*)c_i^{\mathbf{p}}[k] \geq U_j^{'}(r_j^*)c_j^{\mathbf{p}}[k], \quad \forall j \neq i \in \mathcal{M}$$
$$\text{and } r_i^* = \sum_{k \in \mathcal{K}} c_i^{\mathbf{p}}[k] \triangle f x_{ik}^*.$$

The above condition can be also expressed as follows. For a fixed power allocation $\mathbf{p}$ and concave utility functions, a set of D_i^*'s is globally optimal if

$$U_i^{'}(r_i^*)c_i^{\mathbf{p}}[k] \geq U_j^{'}(r_j^*)c_j^{\mathbf{p}}[k], \ \forall k \in D_i^*, \ \forall i, j \in \mathcal{M} \tag{2.27}$$

$$\text{and } r_i^* = \sum_{k \in D_i^*} c_i^{\mathbf{p}}[k] \triangle f. \tag{2.28}$$

Therefore, the variation of the optimality conditions for the continuous frequency case developed in Theorem 2.2 also holds for the discrete frequency case. It is worth noting that the above conditions are only *sufficient* for optimality, and that its *necessity* is lost due to the non-convexity of the achievable data rate region in this case.

We consider two specific cases. First, continuous rate adaptation is used. In this scenario, since the channel fading levels, $H_i[k]$'s, are continuous random variables, the $c_i^{\mathbf{p}}[k]$'s are continuous random variables as well, which implies that $\mathbb{P}\left\{c_i^{\mathbf{p}}[k] = c_{i'}^{\mathbf{p}}[k']\right\} = 0$ for pair $(i,k) \neq (i',k')$. According to (2.27) and (2.28), subcarrier k should be assigned to user m according to the following rule:

$$m(k) = \arg\max_{i \in \mathcal{M}} \{U_i^{'}(r_i^*) \cdot c_i^{\mathbf{p}}[k]\}, \tag{2.29}$$

where $m(k)$ represents that subcarrier k should be assigned to user $m(k)$, and

$$r_i^* = \sum_{k \in D_i^*} c_i^{\mathbf{p}}[k] \triangle f.$$

Another scenario is that linear utility functions are used. For a linear utility function $U_i(r_i)$, its marginal utility function $U_i^{'}(r_i)$ is a constant, which is denoted as $U_i^{'}$. With the linearity, (2.25) becomes

$$U(\mathbf{x}) - U(\mathbf{y}) = \nabla_{\mathbf{x}} U^T(\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{B}.$$

Using the same method, we have that with linear utility functions, a set of D_i^*'s is globally optimal *if and only if*

$$U_i^{'} c_i^{\mathbf{P}}[k] \geq U_j^{'} c_j^{\mathbf{P}}[k], \ \forall k \in D_i^*, \ \forall i, j \in \mathcal{M} \tag{2.30}$$

Therefore, the optimal subcarrier assignment has the following closed form

$$m(k) = \arg\max_{i \in \mathcal{M}} \{U_i^{'} \cdot c_i^{\mathbf{P}}[k]\}. \tag{2.31}$$

2.3.1.2 Sorting-Search Algorithm of Subcarrier Assignment

The utility-based subcarrier assignment optimization belongs to the set of nonlinear combinatorial optimization problems, in which there is no general approach to achieve optimality. In this section, we propose a sorting-search algorithm to seek the optimal subcarrier assignment.

Let us first consider the two-user case, in which each subcarrier in a set of subcarrier indices $\mathcal{A}$ ($\mathcal{A} \subseteq \mathcal{K}$) will be assigned to either user 1 or user 2. For this combinatorial optimization problem, there are $2^{|\mathcal{A}|}$ choices to assign $|\mathcal{A}|$ subcarriers, where $|\mathcal{A}|$ denotes the number of elements in set $\mathcal{A}$. The key idea of sorting-search algorithm is to assume that the conditions (2.27) and (2.28) are both sufficient and necessary. Thus, we have the following rule for an optimal subcarrier assignment: *In the two-user case, if subcarrier i is assigned to user 1, and $\frac{c_2^{\mathbf{P}}[j]}{c_1^{\mathbf{P}}[j]} < \frac{c_2^{\mathbf{P}}[i]}{c_1^{\mathbf{P}}[i]}$, then subcarrier j must be assigned to user 1 as well.* From the above rule, after the $\frac{c_2^{\mathbf{P}}[k]}{c_1^{\mathbf{P}}[k]}$'s for all $k \in \mathcal{A}$ are sorted in an increasing order, there are only $|\mathcal{A}| + 1$ possible assignments that may result in the optimal point, including the two extreme cases: all subcarriers are assigned to user 1 or user 2. In other words, if the conditions (2.27) and (2.28) are both sufficient and necessary, the optimal rate vector should be located on the boundary of the convex hull of the feasible data rate vector set.

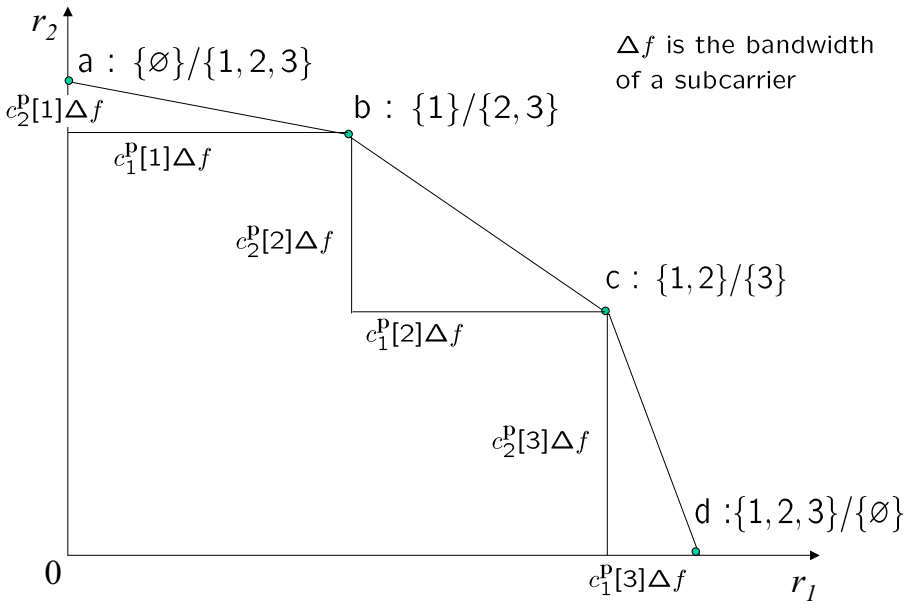


Figure 2.5. An illustration of properties of DSA

For example, if there are three subcarriers for users 1 and 2, and $\frac{c_2^{\mathbf{p}}[1]}{c_1^{\mathbf{p}}[1]} \leq \frac{c_2^{\mathbf{p}}[2]}{c_1^{\mathbf{p}}[2]} \leq \frac{c_2^{\mathbf{p}}[3]}{c_1^{\mathbf{p}}[3]}$, then there are only 4 possible choices for D_1/D_2: $\{\varnothing\}/\{1,2,3\}$, $\{1\}/\{2,3\}$, $\{1,2\}/\{3\}$, $\{1,2,3\}/\{\varnothing\}$. The data rates for users 1 and 2 with those four different subcarrier assignments are shown in Figure 2.5. From the figure, we can see that the slopes of the lines ab, bc, and cd are $-\frac{c_2^{\mathbf{p}}[1]}{c_1^{\mathbf{p}}[1]}$, $-\frac{c_2^{\mathbf{p}}[2]}{c_1^{\mathbf{p}}[2]}$, and $-\frac{c_2^{\mathbf{p}}[3]}{c_1^{\mathbf{p}}[3]}$, respectively. Note that the data rate vectors a, b, c, and d are located on the boundary of the convex hull of the feasible data rate vector set.

The remaining problem is to find out the optimal one among $|\mathcal{A}|+1$ choices. Let T be a threshold that subcarriers satisfying $\frac{c_2^{\mathbf{p}}[k]}{c_1^{\mathbf{p}}[k]} > T$ are assigned to user 2, and the rest to user 1. Therefore, T will determine a subcarrier assignment. From the previous discussion, the optimal T should be close to $\frac{U_1'(r_1)}{U_2'(r_2)}$. With the increase of T, r_1 increases, and r_2 decreases. Because of the concavity of utility functions, $\frac{U_1'(r_1)}{U_2'(r_2)}$ decreases with the increase of T. Clearly, binary search is the best way to arrive at the optimal threshold.

The complexity of this algorithm is very low. The average computational complexity of sorting is about $|\mathcal{A}|\log_2(|\mathcal{A}|)$, and that of binary search is only $\log_2(|\mathcal{A}|)$ [34]. Therefore, the average computational complexity of this algorithm is less than $(K+1)\log_2(K)$.

Algorithm 1 Sorting-Search Subcarrier Assignment for the Two-User Case

sort $\frac{c_2^{\mathbf{P}}[k]}{c_1^{\mathbf{P}}[k]},\ k \in \mathcal{A}$ in increasing order
get thresholds: $T[k],\ k \in \{1 : |\mathcal{A}| + 1\}$ in increasing order
$low = 1; \quad high = |\mathcal{A}| + 1$
while $high - low > 0$ **do**
  $center \leftarrow \lfloor (low + high)/2 \rfloor$
  $T \leftarrow center$
  if $T - \frac{U_1'(r_1)}{U_2'(r_2)} > 0$ **then**
    $high \leftarrow center$
  else
    $low \leftarrow center$
  end if
end while
choose the best T between *low* and *high*

For the M-user case, we can update the subcarrier assignment of every two users iteratively by means of the subcarrier assignment algorithm for the two-user case. Obviously, the computational complexity is nearly $(M-1)^2(K+1)\log_2(K)$, which is still efficient compared to the number of choices of this combinatorial optimization problem, K^M. Moreover, the algorithm is robust to both continuous and discrete rate adaptation.

It should be indicated that the above sorting-search algorithm is in general suboptimal. However, the algorithm will be optimal in each of the following cases.

1. *The utility functions are all linear*: This is because the condition (2.30) is sufficient and necessary for the optimality.

2. *The bandwidth of a subcarrier of the OFDM signal is infinitesimal*: In this case, $\frac{\triangle f}{B} \rightarrow 0$, the feasible data rate region becomes convex. From Section 2.2, this condition leads to the sufficient and necessary condition for optimality

$$\nabla_{\mathbf{x}} U(\mathbf{x})^T (\mathbf{x} - \mathbf{y}) \geq 0 \quad \forall \mathbf{y} \in \mathcal{B}.$$

In practical OFDM systems, $\frac{\triangle f}{B}$ is usually small, and thus the performance of the sorting-search algorithm is nearly optimal in practical situations.

2.3.2 Adaptive Power Allocation Algorithms

We develop algorithms for adaptive power allocation in this section. First, we assume that subcarrier assignment is fixed, and then we extend the developed algorithms to the joint DSA and APA.

2.3.2.1 APA for Fixed Subcarrier Assignment

When only APA is allowed in the system, the subcarrier assignment, D_i for all i, is fixed, and we have

$$\max_{\mathbf{p}} \quad \sum_{i\in\mathcal{M}} U_i(r_i) \tag{2.32}$$

$$\text{subject to} \quad \sum_{k\in\mathcal{K}} p[k] \le \bar{P} \tag{2.33}$$

$$p[k] \ge 0. \tag{2.34}$$

When continuous rate adaptation is used, the optimal power allocation for a fixed subcarrier assignment has the following solution, which comes from Theorem 2.3.

$$p^*[k] = \left[\frac{U_i'(r_i^*)}{\lambda} - \frac{1}{\beta\rho_i[k]}\right]^+ \quad \lambda > 0, \quad k \in D_i \tag{2.35}$$

and λ and r_i^*'s satisfy

$$\sum_{k\in\mathcal{K}} p^*[k] = \bar{P}$$

$$r_i^* = \sum_{k\in D_i} \log_2(1+\beta p^*[k]\rho_i[k])\triangle f.$$

This is actually a utility-based water-filling.

2.3.2.2 Sequential-Linear-Approximation Water-filling Algorithm for Continuous Rate Adaptation

With continuous rate adaptation, the APA optimization is still a non-linear convex programming problem, and (2.35) is both sufficient and necessary for global optimality. When a subcarrier assignment is fixed, the non-linear optimization problem can be approached by a series of linear optimization problems by means of the sequential-linear-approximation algorithm (Frank-Wolfe method) [47], which can be summarized by Algorithm 2. Each

iteration of the algorithm contains two steps. First, we solve an optimization problem with fixed marginal utilities, which is a regular water-filling problem, and then update their marginal utilities using a subgradient method. Intuitively, by solving the group of optimization problems with a linear objective $\sum_{i\in\mathcal{M}} \gamma_i r_i$ subject to the same constraints as those of the original problem, for all possible $\gamma_i \geq 0$, we can trace out the entire boundary of the data rate region.

Algorithm 2 Sequential-Linear-Approximation Water-filling Algorithm for Continuous Rate Adaptation

Iterate until $\sum_{i\in\mathcal{M}} U_i'(r_i^{(n)})(r_i^{(n+1)} - r_i^{(n)}) \leq \epsilon$

1. Get the new power allocation from the linear optimization problem and the corresponding data rates.

$$\begin{aligned} p[k] &\leftarrow \left[\frac{\gamma_{m(k)}^{(n)}}{\lambda} - \frac{1}{\beta\rho_{m(k)}[k]}\right]^+ \text{ for all } k \\ r_i^{(n+1)} &\leftarrow \sum_{k\in D_i} \log_2(1+\beta p[k]\rho_i[k])\triangle f \text{ for all } i, \end{aligned}$$

   where $m(k)$ means that subcarrier k is assigned to user $m(k)$.

2. Update $\gamma_i^{(n)}$ with a positive step-size $\mu \in (0,1)$.

$$\gamma_i^{(n+1)} \leftarrow (1-\mu)\gamma_i^{(n)} + \mu U_i'(r_i^{(n+1)}) \text{ for all } i$$

2.3.2.3 Greedy Power Allocation Algorithm Based on Maximizing Total Utility for Discrete Rate Adaptation

In practice, continuous rate adaptation is infeasible, and there are only several modulation levels. Thus, the optimal power level at each subcarrier for discrete rate adaptation is not continuous either. As a result, the previous water-filling algorithm cannot achieve the optimal power allocation. Therefore, we develop a greedy algorithm for discrete modulation levels.

The key idea of the greedy algorithm is to allocate bits and the corresponding power successively and maximize the utility argument per power in each step of bit loading. Let $f(b)$ be the required power to transmit b bits/sec/Hz, which is usually determined by the

system design. If MQAM is used, according to (2.2), $f(b)$ is given by

$$f(b) = \frac{2^b - 1}{\beta\rho[k]}.$$

In initialization, zero bits are assigned to all subcarriers. During each bit loading iteration, power is allocated at some subcarrier so that the increase of utility per power is maximized. The iteration process will stop when the total transmission power constraint is reached. The greedy power allocation is summarized in Algorithm 3. Note that nonlinear concave utility functions do not increase the algorithm complexity compared to linear utility functions.

Algorithm 3 Greedy Power Allocation Algorithm for Discrete Rate Adaptation

$b_k \leftarrow 0$; $\triangle p_k \leftarrow 0$ for all k
$r_i \leftarrow 0$ for all i
$p_{total} \leftarrow 0$; $\triangle p \leftarrow 0$
while $p_{total} + \triangle p < \bar{P}$ **do**
  $p_{total} \leftarrow p_{total} + \triangle p$
  $\triangle p_k \leftarrow f(b_k + \triangle b_k) - f(b_k)$ for all k,
  where b_k is the current modulation level of subcarrier k, and $\triangle b_k$ is the difference between the next modulation level and the current one for subcarrier k.
  if $p_{total} + \triangle p_k > \bar{P}$ **then**
    $\triangle p_k \leftarrow \infty$
  end if
  $\triangle U_k \leftarrow U_{m(k)}(r_{m(k)} + \triangle b_k) - U_{m(k)}(r_{m(k)})$
  $\hat{k} \leftarrow \arg\max_{k \in \mathcal{K}}(\frac{\triangle U_k}{\triangle p_k})$
  $\triangle p \leftarrow \triangle p_{\hat{k}}$
  if $p_{total} + \triangle p \leq \bar{P}$ **then**
    $b_{\hat{k}} \leftarrow b_{\hat{k}} + \triangle b_{\hat{k}}$
    $r_{m(\hat{k})} \leftarrow r_{m(\hat{k})} + \triangle b_{\hat{k}}$
  end if
end while

Using the following three steps, we can prove that the greedy algorithm results in the *global* optimal bit loading and power allocation with concave utility functions. First, we show that the objective function (2.32) is also concave with respect to the power vector **p**. Then, we check that the feasible region of power allocation vector **p** constrained by (2.33) is a *polymatroid*, which satisfies the normalized, nondecreasing, and submodular properties [21]. Finally, taking advantage of the concavity of objective function and the polymatroid structure of the feasible region, we can demonstrate the optimality of the greedy algorithm according to [21].

2.3.3 Joint Dynamic Subcarrier Assignment and Adaptive Power Allocation

As in Section 2.2, when both power allocation and subcarrier assignment can be changed, the joint DSA and APA optimization problem can be expressed as follows:

$$\max_{D_i, i\in\mathcal{M}, \mathbf{p}} \quad \sum_{i\in\mathcal{M}} U_i(r_i) \tag{2.36}$$

$$\text{subject to} \quad \bigcup_{i\in\mathcal{M}} D_i \subseteq \mathcal{K}, \tag{2.37}$$

$$D_i \bigcap D_j = \varnothing, \quad i \neq j \quad \forall i, j \in \mathcal{M}, \tag{2.38}$$

$$\sum_{k\in\mathcal{K}} p[k] \leq \bar{P} \tag{2.39}$$

$$p[k] \geq 0. \tag{2.40}$$

Obviously, the optimal resource allocation (optimal rate vector) must simultaneously satisfy the conditions for the DSA-only and APA-only problems. Similar to the discussion in Section 2.3.2.2, for concave functions, the algorithm for the joint DSA and APA using continuous rate adaptation is a combination of iterative subcarrier assignment, power allocation, and the update of marginal utility, which is summarized in Algorithm 4. For those concave utility functions, using this algorithm with an appropriate update-step μ, we can find a global maximum. The computational complexity of the subcarrier assignment is only $\mathcal{O}(MK)$. For discrete rate adaptation, we can iteratively use the sorting-search DSA and the greedy APA algorithms.

2.3.4 Algorithm Modification for Nonconcave Utility Functions

Utility functions depend on the type of applications and are not always concave. For instance, it is demonstrated in [29] that the utility function for best-effort applications is

$$U(r) = [0.16 + 0.8\ln(r - 0.3)]^+, \tag{2.41}$$

where the r is in units of kbps. For a more general use, we express the utility function as

$$U(r) = \begin{cases} a + b\ln(r - c) & r \geq r_{thr} \\ 0 & 0 \leq r < r_{thr} \end{cases} \tag{2.42}$$

where $b > 0$ and $a = -b\ln(r_{thr} - c)$ is a threshold. Even though the above utility function (2.41) is not exactly concave over $[0, +\infty)$, it is strictly concave and differentiable when

Algorithm 4 Joint DSA and APA with Continuous Rate Adaptation

Iterate until $\sum_{i\in\mathcal{M}} U_i'(r_i^{(n)})(r_i^{(n+1)} - r_i^{(n)}) \leq \epsilon$

1. Get the new subcarrier assignment, according to the condition (2.30), using

$$m(k) \leftarrow \arg\max_{i\in\mathcal{M}} \{\gamma_i^{(n)} c_i^{\mathbf{p}}[k]\} \quad \text{for all } k$$

2. Get the new power allocation from the linear optimization problem and the corresponding data rates.

$$\begin{aligned} p[k] &\leftarrow \left[\frac{\gamma_{m(k)}^{(n)}}{\lambda} - \frac{1}{\beta\rho_{m(k)}[k]}\right]^+ \quad \text{for all } k \\ r_i^{(n+1)} &\leftarrow \sum_{k\in D_i} \log_2(1+\beta p[k]\rho_i[k])\triangle f \quad \text{for all } i \end{aligned}$$

3. Update $\gamma_i^{(n)}$ with a positive step-size $\mu \in (0,1)$.

$$\gamma_i^{(n+1)} \leftarrow (1-\mu)\gamma_i^{(n)} + \mu U_i'(r_i^{(n+1)}) \quad \text{for all } i$$

the data rate is above a threshold. For this utility function (2.41), the threshold, $r_{thr} = 1.119$ kbps, is very small. As a result, the non-concavity of this function may not significantly affect the solution of the optimization problem, especially in the case of high SNR. However, the non-concavity sometimes does affect the solution; therefore, we will propose an approach to deal with the non-concavity problem.

Finding the global optimum of a non-convex optimization problem is in general very difficult. An intuition can be obtained from (2.42); this utility function implies the need of admission control. r_{thr} is actually the threshold for admission control. Our solution includes the following two steps:

- Modifying this utility function to $\tilde{U}(r)$ as follows:

$$\tilde{U}(r) = \begin{cases} U(r), & r \geq r_{thr}, \\ U'(r_{thr})(r - r_{thr}), & 0 \leq r < r_{thr}, \end{cases} \tag{2.43}$$

  which is concave over $[0, +\infty)$, and a global maximum for the modified utility function can be obtained by using the previous algorithms. Note that the modification is suitable for any utility function that is concave over $[r_{thr}, +\infty)$.

- Using admission control, shown in Figure 2.6, the solution obtained from the modified utility function $\tilde{U}(r)$ can be corrected to that of the original utility function $U(r)$.

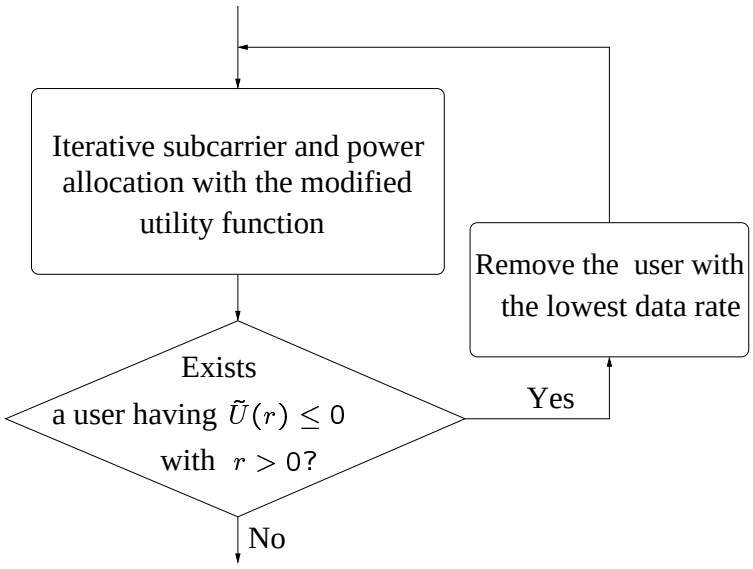


Figure 2.6. Modified dynamic resource allocation algorithm

2.4 Cross-Layer Optimization Based on Utility Functions With Respect to Average Date Rates

All of the forementioned resource allocation algorithms underlying maximizing the aggregate utility just consider the instantaneous channel conditions, fairness, and efficiency. In reality, however, users mainly care about the average data rate during a certain period of time, not the instantaneous one. In this section, we investigate the impact of time diversity on the performance of the cross-layer optimization. We start with a general case and then study the asymptotic performance.

The average data rate $\bar{r}_i[n]$ of each user at time n can be expressed by using an exponentially weighted low-pass time window as

$$\bar{r}_i[n] = (1 - \rho_w)\bar{r}_i[n-1] + \rho_w r_i[n]. \tag{2.44}$$

where $r_i[n]$ is the instantaneous data rate of user i at time n. $\rho_w = \dfrac{T_s}{T_w}$, where T_s is the slot length, and T_w is the length of the window. Therefore, the optimization problem should be expressed as maximizing the total utility with respect to the average data rates, $\bar{r}_i[n]$'s,

that is,

$$\max_{\mathbf{r}[n]\in\mathcal{C}_\pi(\mathbf{H})} \sum_{i\in\mathcal{M}} U_i(\bar{r}_i[n]) \tag{2.45}$$

where $\mathbf{r}[n]$ is the data rate vector $[r_1[n], r_2[n], \cdots, r_M[n]]^T$, and $\mathcal{C}_\pi(\mathbf{H})$ is the instantaneous feasible data rate region at time n, determined by the current channel states $\mathbf{H}$, which is given by

$$\mathbf{H} = (H_1[1], \cdots, H_1[K], \cdots, H_M[1], \cdots, H_M[K]),$$

and the allocation constraints of a certain resource allocation policy π, such as DSA, APA, as well as joint DSA and APA.

Since $\bar{r}_i[n]$ is a function of $r_i[n]$ in (2.44), the optimization problem (2.45) can be rewritten as

$$\max_{\mathbf{r}[n]\in\mathcal{C}_\pi(\mathbf{H})} \sum_{i\in\mathcal{M}} V_i(r_i[n]) \tag{2.46}$$

where $V_i(r_i[n]) = U_i\left((1-\rho_w)\bar{r}_i[n-1] + \rho_w r_i[n]\right)$.

The above problem can be regarded as an optimization based on utility functions $V_i(\cdot)$'s with respect to the instantaneous data rates $r_i[n]$'s as well. The marginal utility function is given by

$$\frac{\partial}{\partial r_i[n]} V_i(r_i[n]) = \rho_w U_i'(r_i)\big|_{r_i=(1-\rho_w)\bar{r}_i[n-1]+\rho_w r_i[n]},$$

Therefore, all previous algorithms work well as long as $U_i(r_i)$ and $U_i^{'}(r_i)$ are replaced by $U_i(\bar{r}_i[n])$ and $\rho_w U_i^{'}(\bar{r}_i[n])$, respectively. The computational complexity of maximizing the total utility function with respect to the average data rates is the same as that of optimization with respect to the instantaneous data rates.

Without a time window, the optimization problem must guarantee fairness in each slot period. However, when the time window is used, the fairness requirement is relaxed to a time-window length. This provides more flexibility to improve the spectral efficiency. On the other hand, the current resource allocation is related to the previous ones. If one user has a higher average date rate, his priority is set to be lower. Hence, the use of a time window may enhance fairness as well. The length of the time window should be longer than

the correlation time of the channel in order to get more time diversity. But if it is too long, the utility function cannot capture the short-term preference of users.

If ρ_w is small enough, the computational complexity of the optimization problem can be reduced furthermore. With a small ρ_w, we have

$$\frac{\partial U_i(\bar{r}_i[n])}{\partial r_i[n]} \approx \rho_w U_i'(r_i)\big|_{r_i=\bar{r}_{i[n-1]}},$$

which means that the current marginal utility values are totally determined by the previous resource allocation. With one-order Taylor formula, it follows that

$$\begin{aligned}&\sum_{i\in\mathcal{M}} U_i(\bar{r}_i[n]) - \sum_{i\in\mathcal{M}} U_i(\bar{r}_i[n-1])\\ &\approx \sum_{i\in\mathcal{M}} U_i'(\bar{r}_i[n-1])(\rho_w r_i[n] - \rho_w \bar{r}_i[n-1])\end{aligned} \tag{2.47}$$

Since all $\bar{r}_i[n-1]$'s are fixed at time n, the optimization problem becomes the one with a linear objective function as follows,

$$\max_{\mathbf{r}[n]\in\mathcal{C}_\pi(\mathbf{H})} \sum_{i\in\mathcal{M}} U_i'(\bar{r}_i[n-1]) r_i[n], \tag{2.48}$$

which maximizes the sum of weighted rates. The weights are adaptively controlled by the marginal utility with respect to the current average rates.

The linear objective function greatly simplifies the corresponding algorithms. In particular, for DSA, we have the following closed form according to (2.31),

$$m(k,n) = \arg\max_{i\in\mathcal{M}} \{U_i^{'}(\bar{r}_i[n-1]) \cdot c_i^{\mathbf{P}}[k,n]\}, \tag{2.49}$$

where $m(k,n)$ represents that subcarrier k is assigned to user $m(k,n)$ at time n, and $c_i^{\mathbf{P}}[k,n]$ denotes the achievable transmission efficiency of subcarrier k at time n. Its complexity is only $M \cdot K$. If there is only one carrier (single-carrier system), and if $U_i(\bar{r}_i[n]) = \ln(\bar{r}_i[n])$, and (2.49) is simplified as

$$m(n) = \arg\max_{i\in\mathcal{M}} \{\frac{c_i^{\mathbf{P}}[n]}{\bar{r}_i[n-1]}\},$$

which is just the proportional fair scheduling proposed for CDMA systems in [76]. Therefore, the utility-based resource allocation we presented here is a general framework for allocating multiuser shared resources. For APA or joint DSA and APA, iteration is still needed, but the linear objective function offers fast convergence.

2.5 Efficiency and Fairness

Both efficiency and fairness issues are very important for resource allocation in wireless networks. An allocation scheme is said to be *efficient* if there is no other scheme that would simultaneously benefit someone and harm nobody in terms of their utilities. Therefore, the utility-based optimization is obviously efficient. Note that it differs from the *spectral efficiency* that is measured in terms of the total throughput over the bandwidth. Clearly, the maximum spectral efficiency is achieved by using a utility function $U_i(r_i) = r_i$ for all i.

With the channel knowledge for each user at the base station, the DSA scheme tends to assign subcarriers to users with a better SNR at the corresponding subcarriers, thereby having high spectral efficiency. It is obvious from (2.13) that the utility-based DSA penalizes the users with poor channel conditions.

When $U_i(r_i) = r_i$, $U_i^{'}(r_i) = 1$. In this case each subcarrier is assigned to the user with the best channel conditions among all users; therefore, the system can obtain the largest multiuser diversity with respect to spectral efficiency. Although the multiuser diversity is similar to the traditional selection diversity, its diversity gain results from the number of users, rather than from the number of antennas.

Fairness requires a fair share of bandwidth among competing users and protection from aggressive connections. Two representative types of fairness are *proportional* fairness [30] and *max-min* fairness [10]. Proportional fairness provides each connection a priority inversely proportional to its data rate. A vector of rates $\mathbf{r} \in \mathcal{C}$ is said to be *proportionally fair* if for any other feasible rate vector $\mathbf{r}' \in \mathcal{C}$, the aggregate of proportional changes is zero or negative:

$$\sum_{i=1}^{M} \frac{r_i' - r_i}{r_i} \leq 0. \tag{2.50}$$

For a concave utility function $U(\mathbf{r})$ and a convex set $\mathcal{C}_\pi$, $\mathbf{r}$ is optimal if and only if

$$\nabla U(\mathbf{r})^T(\mathbf{r}' - \mathbf{r}) \leq 0 \quad \text{for all } \mathbf{r}' \in \mathcal{C}_\pi. \tag{2.51}$$

where $\nabla U(\mathbf{r}) = [U_1'(r_1), U_2'(r_2), \cdots, U_M'(r_M)]^T$. When the logarithmic utility function, $U(r) = \ln(r)$, is used, (2.51) is identical to (2.50). Therefore, the logarithmic utility function is associated with the proportional fairness for the utility-based optimization.

A data rate vector $\mathbf{r}$ is *max-min fair* if for each $m \in \mathcal{M}$, r_m cannot be increased without decreasing r_i form some i for which $r_i < r_m$. Obviously, max-min fairness has a strict fairness criterion since lower rates can get an absolute priority.

Consider a family of utility functions that is expressed as

$$U(r) = -\frac{r^{-\alpha}}{\alpha}, \quad \alpha > 0. \tag{2.52}$$

Obviously, the parameter α determines the degree of fairness. As α increases, the fairness of the corresponding utility function becomes stricter and stricter. When $\alpha \to \infty$, it turns out to be the max-min fairness.

It can be also seen from (2.13) that increasing utility functions encourage the users having good channel conditions, and decreasing marginal utility functions assign a high priority to the users with a low data rate. Therefore, utility-based resource allocation can guarantee both efficiency and fairness.

2.5.1 Fairness of "Extreme OFDM" Using Utility Functions With Respect to Instantaneous Data Rates

Since the instantaneous capacity is convex in the "extreme OFDM" system in Section 2.2, in which the number of subcarrier is assumed to be infinite, utility-based optimization related to instantaneous data rates in Section 2.2 can maintain a fairness defined as (2.51) with respect to the instantaneous capacity region.

2.5.2 Fairness of "Practical OFDM" Using Utility Functions With Respect to Average Data Rates

In practical OFDM systems, in which the number of subcarrier is finite, the instantaneous capacity region is not convex anymore. Thus, we focus on the steady state of the system and the fairness related to the long-term average data rate region.

The scheduling algorithm (2.49) is assumed to be used. We consider the situation in steady state when the window size $T_w \to \infty$. It is assumed that the channel processes $\mathbf{H}$ are ergodic. We denote by $\tilde{r}_i$ the limit data rate of user i; due to the ergodicity of $\mathbf{H}$, it

follows that

$$\begin{aligned}\tilde{r}_i &= \lim_{n\to\infty} \bar{r}_i[n] \\ &= \mathbb{E}\{r_i\},\end{aligned}$$

Let $\tilde{\mathcal{C}}_\pi$ be the long-term *average data rate region* under the allocation constraints of the policy π, which consists of the average data rate vectors obtained by all possible *stationary* resource allocation schemes.

It is easy to prove that $\tilde{\mathcal{C}}_\pi$ is a convex set; that is, $\forall\ \tilde{\mathbf{r}}^{(1)}, \tilde{\mathbf{r}}^{(2)} \in \tilde{\mathcal{C}}_\pi, \alpha \in [0,1]$, we will show that $\alpha\tilde{\mathbf{r}}^{(1)} + (1-\alpha)\tilde{\mathbf{r}}^{(2)} \in \tilde{\mathcal{C}}_\pi$. According to the definition of $\tilde{\mathcal{C}}_\pi$, there must exist such a resource allocation scheme $F^{(1)}$ that $\tilde{\mathbf{r}}^{(1)} = \mathbb{E}\{F^{(1)}(\mathbf{H})\}$, where $F^{(1)}(\mathbf{H})$ is the data rate vector of user i under the channel states $\mathbf{H}$ when a resource allocation scheme $F^{(1)}$ is employed. Likely, $\tilde{\mathbf{r}}^{(2)}$ results from another resource allocation scheme $F^{(2)}$ so that $\tilde{\mathbf{r}}^{(2)} = \mathbb{E}\{F^{(2)}(\mathbf{H})\}$. We can construct such a new scheme F that under the channel states $\mathbf{H}$,

$$F = \begin{cases} F^{(1)} & \xi = 1 \\ F^{(2)} & \xi = 0 \end{cases},$$

where ξ is a binary random variable with $P\{\xi = 1\} = \alpha$. It follows that

$$\begin{aligned}\tilde{\mathbf{r}} &= \mathbb{E}\{F(\mathbf{H})\} \\ &= \mathbb{E}\{\xi F^{(1)}(\mathbf{H}) + (1-\xi)F^{(2)}(\mathbf{H})\} \\ &= P\{\xi = 1\}\mathbb{E}\{F^{(1)}(\mathbf{H})\} + (1 - P\{\xi = 1\})\mathbb{E}\{F^{(2)}(\mathbf{H})\} \\ &= \alpha\tilde{\mathbf{r}}^{(1)} + (1-\alpha)\tilde{\mathbf{r}}^{(2)}\end{aligned}$$

The data rate vector $\tilde{\mathbf{r}}$ with respect to the scheme F lies in $\tilde{\mathcal{C}}_\pi$; therefore, $\alpha\tilde{\mathbf{r}}^{(1)} + (1-\alpha)\tilde{\mathbf{r}}^{(2)} \in \tilde{\mathcal{C}}_\pi$.

The optimization problem (2.48) in steady state can be expressed as

$$\max_{\mathbf{r}\in\mathcal{C}_\pi(\mathbf{H})} \sum_{i\in\mathcal{M}} U_i'(\tilde{r}_i) r_i, \tag{2.53}$$

where $\tilde{\mathbf{r}}$ is the steady-state data rate vector. The data rate vector $\mathbf{r}^*$ under the channel conditions $\mathbf{H}$ with respect to the scheme (2.53) is given by

$$\mathbf{r}^* = \arg \max_{\mathbf{r}\in\mathcal{C}_\pi(\mathbf{H})} \sum_{i\in\mathcal{M}} U_i'(\tilde{r}_i) r_i. \tag{2.54}$$

Due to being in steady state, $\mathbb{E}\{\mathbf{r}^*\} = \tilde{\mathbf{r}}$. Obviously, with any other scheme F, it follows that $\mathbf{r}' = F(\mathbf{H})$, and

$$\nabla U(\tilde{\mathbf{r}})^T(\mathbf{r}' - \mathbf{r}^*) \leq 0, \quad \mathbf{r}' \in \mathcal{C}_\pi(\mathbf{H}), \tag{2.55}$$

where, $\nabla U(\tilde{\mathbf{r}}) = [U_1'(\tilde{r}_1), U_2'(\tilde{r}_2), \cdots, U_M'(\tilde{r}_M)]^T$. Taking expectation on both sides, we have

$$\nabla U(\tilde{\mathbf{r}})^T(\tilde{\mathbf{r}}' - \tilde{\mathbf{r}}) \leq 0, \quad \tilde{\mathbf{r}}' \in \tilde{\mathcal{C}}_\pi. \tag{2.56}$$

Due to the convexity of the feasible average rate region $\tilde{\mathcal{C}}_\pi$ and the concavity of utility functions $U_i(r)$'s, the condition (2.56) is sufficient and necessary for the optimality of the following problem [56]

$$\max_{\tilde{\mathbf{r}} \in \tilde{\mathcal{C}}_\pi} \sum_{i \in \mathcal{M}} U_i(\tilde{r}_i). \tag{2.57}$$

Therefore, when $\rho_w \to 0$, the optimization problem in the instantaneous rate region (2.48) can achieve the optimality of the optimization problem with respect to the long-term average data rates in the average rate region (2.57). In this scenario, the properties of efficiency and fairness that utility functions offer are all concerned with long-term average data rates. For instance, if $U_i(\tilde{r}_i) = \ln(\tilde{r}_i)$, then $U_i'(\tilde{r}_i) = 1/\tilde{r}_i$. It follows from (2.56) that

$$\sum_{i \in \mathcal{M}} \frac{\tilde{r}_i' - \tilde{r}_i}{\tilde{r}_i} \leq 0, \quad \text{for all } \tilde{\mathbf{r}}' \in \tilde{\mathcal{C}}_\pi,$$

in which the long-term average data rate vector $\tilde{\mathbf{r}}$ is proportionally fair.

2.6 Simulation Results

In this section, we present simulation results to illustrate the performance of the various resource allocation approaches developed in this chapter. In our simulation, the channel is assumed to have a *bad-urban* (BU) delay profile [70] and suffer from shadowing with 8.0 dB standard deviation. Let the acceptable BER be 10^{-6} for rate adaptation. The bandwidth of each subcarrier is 10 kHz, and the utility function in (2.41) is used. To be able to compare those results properly, we set the average bandwidth per user, B/M, to be 80 kHz and show the average total utility per 80 kHz in simulation results.

At first, we assume the distances from the base station to all users are identical and compare various resource allocation schemes without a time window. Figure 2.7 shows some numerical results for different resource allocation schemes. Here continuous rate adaptation is used for all schemes. The *fixed subcarrier assignment* (FSA) results in the same performance when the number of users changes, while DSA offers significant multiuser diversity, which increases with the number of users. However, in the continuous rate adaptation case, the joint DSA and APA only leads to a very small improvement compared to DSA. The contribution of the APA is limited in this case as well.

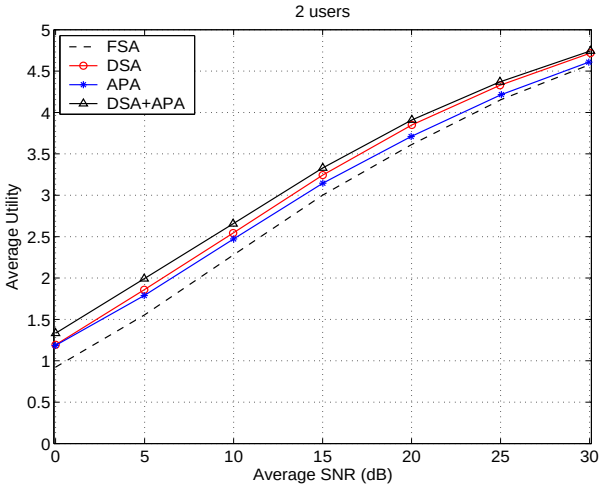
Figure 2.8 shows the performance of different adaptive resource allocation policies with discrete rate adaptation. The variable MQAM with modulation levels $\{0, 2, 4, 6, \cdots\}$ is assumed to be employed. The improvement from the DSA is similar to that in the continuous rate adaptation case. However, the contribution of the APA is significant in sharp contrast to that of continuous rate adaptation. Besides, the DSA in conjunction with the APA is able to substantially improve the network performance even in the two-user case. For example, to achieve an average utility of 3, the gain from the joint DSA and APA is about 8 dB for the two-user case, and it increases to around 11 dB for the 16-user case.

Next, we evaluate fairness and spectral efficiency in the scenario when the distances between users and the base station are different. The path loss is modeled by

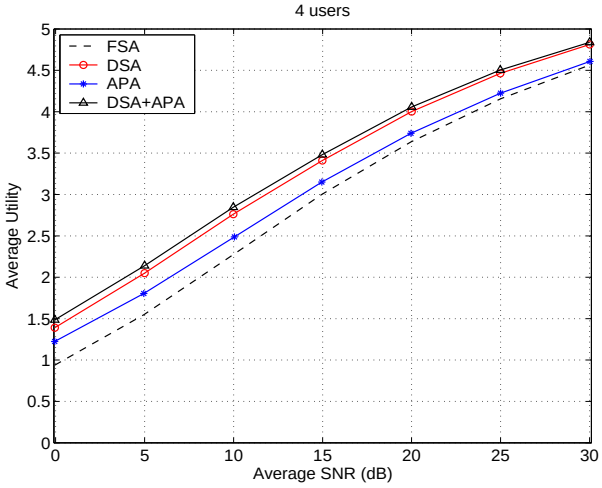
$$PL(d) = 128.1 + 37.6 \log_{10} d \quad [dB]$$

where d (km) is the distance between a user and the base station. Each user is assumed to be stationary or slowly moving so that the maximum Doppler shift is 5 Hz; as a result, their path loss and shadowing values are fixed during the simulation. In this simulation, the number of users is 8. We sort the 8 users according to their distances to the base station. The path loss difference between the users closest to and fartherest from the base station is about 18 dB in the simulation.

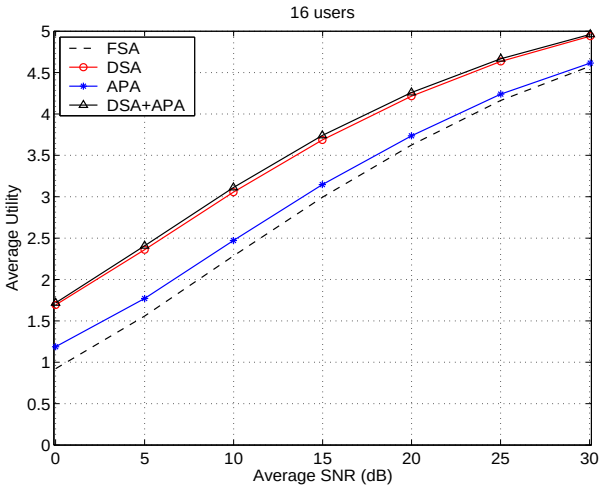
Figures 2.9 and 2.10 show the average throughput and average of each user with various resource allocation policies when continuous and discrete rate adaptation techniques are deployed, respectively. It is clear that although each user gets the same bandwidth and



(a) 2 users

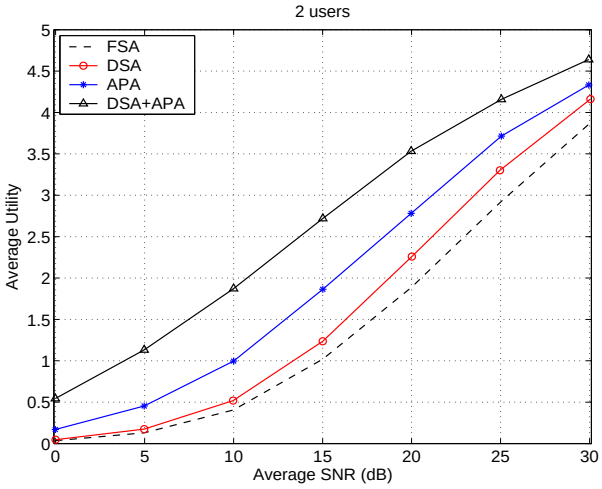


(b) 4 users

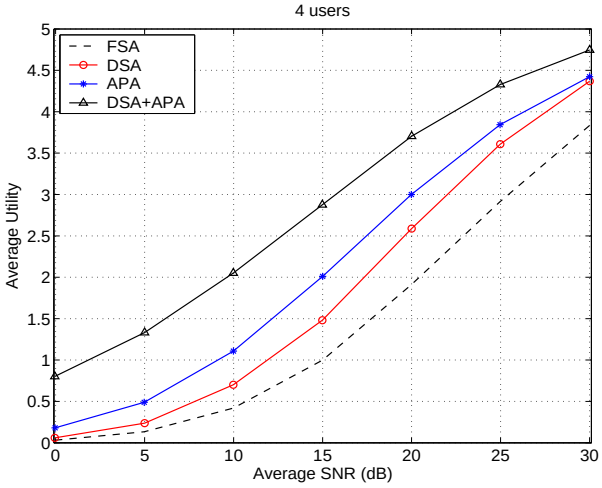


(c) 16 users

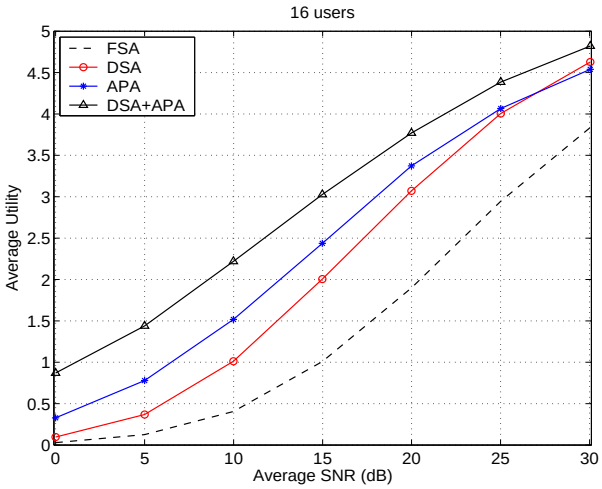
Figure 2.7. Average user utility versus SNR for OFDM wireless network with different resource allocation schemes



(a) 2 users



(b) 4 users



(c) 16 users

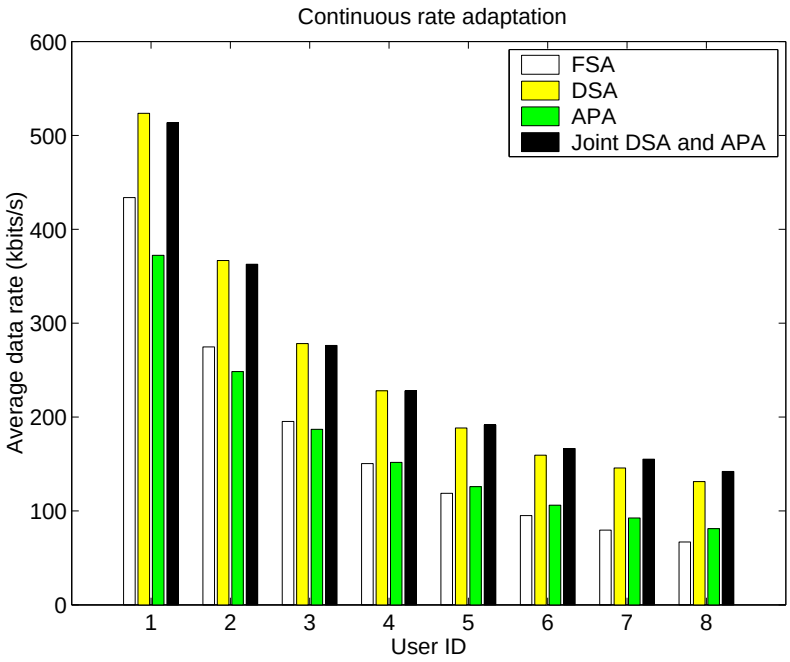
Figure 2.8. Average user utility versus SNR by using discrete rate adaptation and different resource allocation schemes

power, different path loss values result in different data rates. It can be seen from both figures that all utility-based resource allocation policies can increase the total throughput. Furthermore, the poorer the channel conditions, the greater improvement in throughput; hence all utility-based resource allocation schemes can provide fairer services than the FSA. Figure 2.9 also confirms that using the DSA can provide the similar performance as the joint DSA and APA for continuous rate adaptation. For discrete rate adaptation, however, Figure 2.10 shows the significant improvement of the joint DSA and APA in offering efficient and fair allocation.

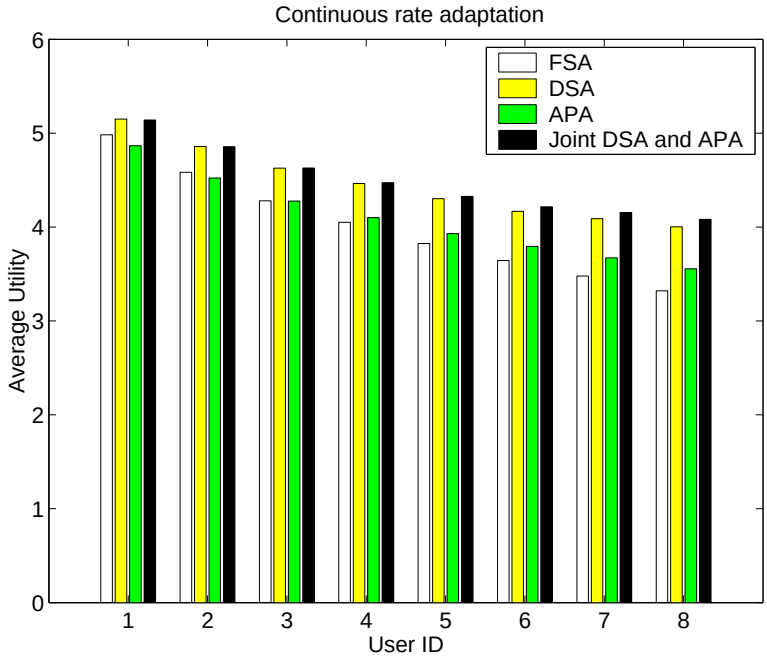
Figure 2.11 demonstrates the performance of the DSA and the joint DSA and APA over time windows with different window lengths when discrete rate adaptation is employed. For the DSA, the time window helps to enhance the fairness of resource allocation. On the other hand, for the joint DSA and APA, a time window can further improve the average throughput of each user. However, the complexity of implementation time windows is negligible.

2.7 Summary

In this chapter, we have presented utility-based cross-layer optimization for OFDM-based wireless networks. The utility is used here to build a bridge between the physical and MAC layers and to balance the efficiency and fairness of resource allocation. In particular, we have investigated the necessary and sufficient conditions for finding an optimum for the DSA, APA, and joint DSA and APA schemes when instantaneous-rate-based utility functions are used and the number of subcarriers is assumed to be infinite. Based on the theoretical framework, we have developed a variety of efficient algorithms, including the sorting-search DSA, the greedy bit-loading and power allocation, and the objective aggregation algorithms for practical OFDM systems. We have also modified the algorithms for non-concave utility functions. The use of average-rate-based utility functions is very suitable to best-effort traffic. A low-pass time filter resulting from average-rate-based utility functions can easily be incorporated into all algorithms to exploit time diversity,. The extensive computer simulation results demonstrate the significant performance gain for the

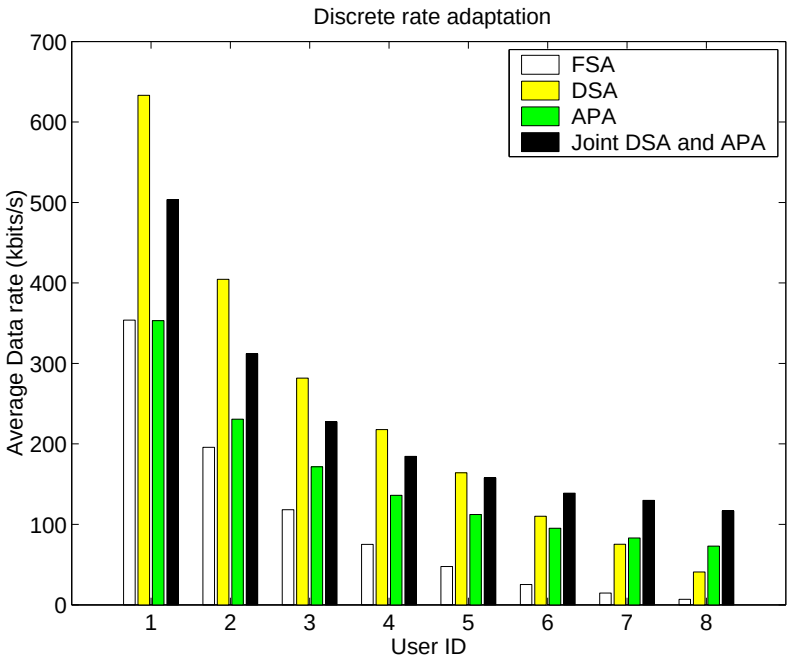


(a) Average throughput performance

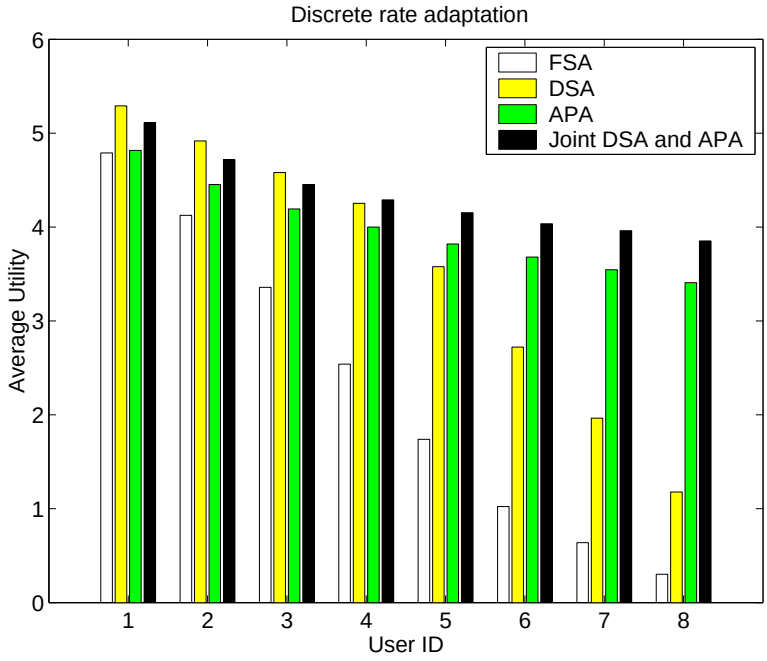


(b) Average utility performance

Figure 2.9. Average performance of various resource allocation schemes with continuous rate adaptation

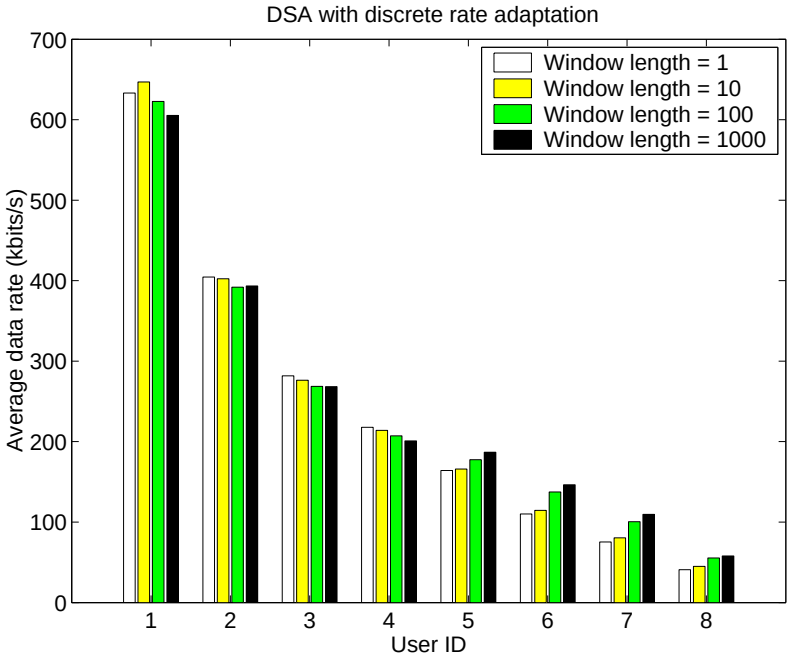


(a) Average throughput performance

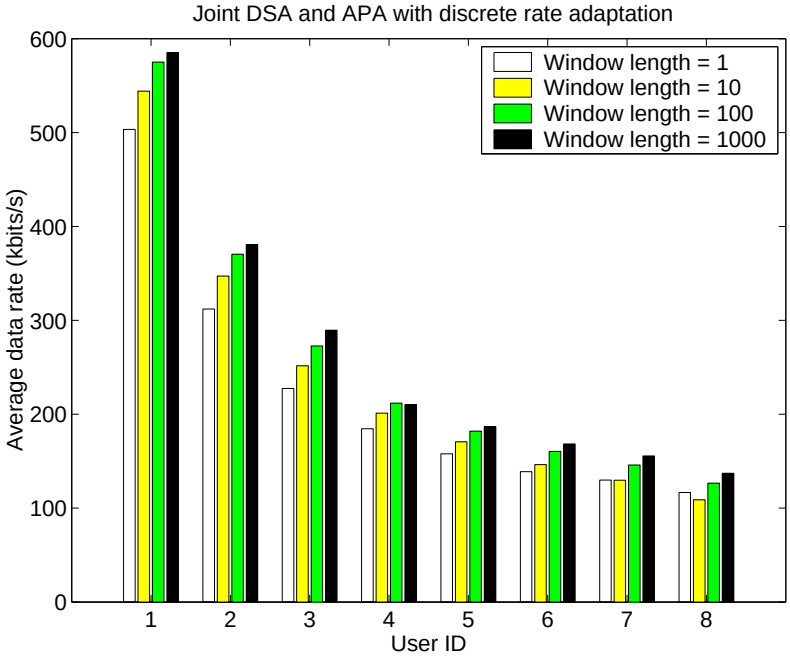


(b) Average utility performance

Figure 2.10. Average performance of various resource allocation schemes with discrete rate adaptation



(a) The performance of DSA over time window



(b) The performance of joint DSA and APA over time window

Figure 2.11. Performance of addition of time window

developed algorithms. In the next Chapter, we will focus on using utility-based optimization for delay-sensitive traffic.

CHAPTER 3

JOINT CHANNEL- AND QUEUE-AWARE MULTICARRIER SCHEDULING USING DELAY-BASED UTILITY FUNCTIONS

The relationship between rate-based utility functions and fairness in wireless networks has been shown in Chapter 2. Rate-based scheduling schemes, which apply the CSI and rate-based utility functions, do not take traffic burstiness into account. In this chapter, utility functions with respect to average delays is used for designing channel- and queue-aware scheduling, which is highly advantageous to data transmission with a low latency requirement.

This chapter is organized as follows. In Section 3.1, we introduce the background and motivations of this work. In Section 3.2, we briefly introduce how to extend scheduling schemes existing in single-carrier systems into the corresponding multichannel scheduling schemes. In Section 3.3, we develop the MDU scheduling based on maximizing the total utility in terms of average waiting time. In Section 3.4, we state the maximum stability region and develop the results regarding stability. In Section 3.6, we propose using delay transmit diversity and adaptive power allocation to further improve the system performance. Finally, in Section 3.7, we compare several multicarrier scheduling schemes using simulation.

3.1 Introduction

It is increasingly clear that most information traffic would be delivered based on IP networks because of the efficient bandwidth use and the low-cost infrastructure construction. Thus, the queue state information, such as queue length and packet delay, which is a reflection of traffic burstiness, should be utilized in scheduling packets. On the other hand, since the queue state information is tightly connected with QoS, wisely controlling queues is one of the

most effective ways for QoS provisioning. As compared to channel-aware scheduling, joint channel- and queue-aware scheduling would be more beneficial to wireless resource allocation and QoS provisioning. *Modified largest weighted delay first* (M-LWDF) and *exponential* (EXP) scheduling rules have been proposed for CDMA downlink transmission in [8, 63], respectively. Neither rules require statistical information about arrival traffic and wireless channels. The stability properties of the M-LWDF and the EXP scheduling rules over time-varying channels have been also studied by using the fluid limit technique in [8, 63], respectively. Other work on packet scheduling with emphasis on queueing system stability can be found in [20, 45, 49, 71, 72].

In this chapter, we investigate joint channel- and queue-aware scheduling in OFDM-based networks with emphasis on designing joint channel- and queue-aware scheduling schemes for multicarrier networks. It should be indicated that the scheduling design for multicarrier networks is not just a simple extension of existing scheduling approaches in single-carrier networks. First, multicarrier networks have nice granularity for resource allocation since the whole bandwidth is divided into many subchannels. Second, multicarrier scheduling actually works in a parallel fashion. Unlike in single-carrier networks, multiple users can be served simultaneously in multicarrier networks; thus, from a queueing point of view, there are multiple servers in multicarrier scheduling.

3.2 Extending Scheduling Rules in Single-Carrier Networks into OFDM Networks

In this section, DSA is used, but power allocation is fixed. Some existing scheduling schemes exploiting multiuser diversity in single-carrier networks can be directly extended to multicarrier networks. In dynamically assigning subcarriers, we usually need to solve the optimization problem expressed as follows:

$$\max_{D_i^{(n)}, i\in\mathcal{A}^n} \sum_{i\in\mathcal{A}^n} w_i[n] r_i[n] \tag{3.1}$$

$$\text{subject to} \bigcup_{i\in\mathcal{A}^n} D_i^{(n)} \subseteq \mathcal{K}, \tag{3.2}$$

$$D_i^{(n)} \bigcap D_j^{(n)} = \varnothing, \quad i \neq j \quad \forall i, j \in \mathcal{A}^n, \tag{3.3}$$

where $\mathcal{A}^n = \{i : Q_i[n] > 0\}$ is the set in which each queue is not empty at time slot n, and the optimization objective is to maximize the sum weighted data rate with the weights $w_1[n], w_2[n], \cdots, w_M[n]$. In Section 2.3.1, the optimal assignment for the above problem is derived as

$$m(k,n) = \arg\max_{i \in \mathcal{A}^n} \{w_i[n] c_i[k,n]\}, \tag{3.4}$$

where $m(k,n)$ ($m(k,n) \in \mathcal{A}^n$) represents subcarrier k to be assigned to user $m(k,n)$ at time n. This result is very useful to design scheduling approaches or to extend some scheduling rules in the single-carrier case to the OFDM scenario.

3.2.1 Max-Sum-Capacity (MSC) Rule

The MSC rule is a channel-aware scheduling scheme that maximizes the total throughput in the system. Thus, the optimization problem can be expressed as (3.1)-(3.3) with $w_i[n] = 1$, for all i. Clearly, the MSC rule is given by

$$m(k,n) = \arg\max_{i \in \mathcal{A}^n} \{c_i[k,n]\}. \tag{3.5}$$

Although the MSC rule makes the most efficient use of the bandwidth, it can lead to unfairness and instability, especially for nonsymmetrical channel conditions and nonuniform traffic patterns.

3.2.2 Proportional Fair (PF) Scheduling

The PF scheduling is a channel-aware scheduling rule aiming to maximize $\sum_i \ln(\bar{r}_i[n])$, where $\bar{r}_i[n]$ is the average data rate for user i. The scheduling rule in multicarrier networks is obtained in [67, 76] also in Chapter 2 as

$$m(k,n) = \arg\max_{i \in \mathcal{A}^n} \left\{ \frac{c_i[k,n]}{\bar{r}_i[n]} \right\}. \tag{3.6}$$

Since ρ_w is very small, $\bar{r}_i[n] \approx \bar{r}_i[n-1]$. Although this DSA algorithm guarantees the proportional fairness [67], it is not throughput-optimal [1] [7]. The PF scheduling is suitable to best-effort traffic, which has no specific QoS requirements.

[1] A scheduling algorithm is called *throughput-optimal* if it stabilizes a queueing system in which stability is feasible at all to do with any algorithms [8] .

3.2.3 Modified Largest Weighted Delay First (M-LWDF) Rule

In [8], the M-LWDF scheme is proposed for single-carrier CDMA networks with a shared downlink channel. From an optimization point of view, the M-LWDF intends to maximize $\sum_i \frac{T_{\mathrm{HOL},i}[n]}{\bar{r}_i[n]} r_i[n]$, where $T_{\mathrm{HOL},i}$ is the delay of the *head-of-line* (HOL) packet of user i. Using the result (3.4), we have the multichannel version of M-LWDF as

$$m(k,n) = \arg\max_{i\in\mathcal{A}^n} \left\{ \frac{c_i[k,n]}{\bar{r}_i[n]} T_{\mathrm{HOL},i[n]} \right\}.$$

3.2.4 Exponential (EXP) Rule

The EXP scheduling rule is also designed for single-carrier CDMA networks with a shared downlink channel [63]. The structure of the EXP rule is very similar to the M-LWDF, but with different weights. The multichannel version of EXP rule can be expressed as

$$m(k,n) = \arg\max_{i\in\mathcal{A}^n} \left\{ \frac{c_i[k,n]}{\bar{r}_i[n]} \exp\left(\frac{T_{\mathrm{HOL},i}[n]}{1+\sqrt{\overline{T}_{\mathrm{HOL}}[n]}} \right) \right\},$$

where $\overline{T}_{\mathrm{HOL}}[n] = \frac{1}{|\mathcal{A}^n|} \sum_{i\in\mathcal{A}^n} T_{\mathrm{HOL},i}[n]$.

The M-LWDF and EXP rules have been proven to be throughput-optimal in single-carrier networks [8, 63]. With a few modifications, the proofs are valid in OFDM networks. Both scheduling rules are proposed for delay-sensitive traffic.

3.3 Max-Delay-Utility (MDU) Scheduling

Designing channel-aware-only scheduling is usually tractable. It is shown in Chapter 2 that most channel-aware-only scheduling schemes can be derived by maximizing the sum of specific utility functions with respect to data rates. However, there are two difficulties in designing joint channel- and queue-aware scheduling. First, it is hard to formulate the desired optimization goals related to the QoS requirements such as average waiting time, delay violation probability, etc. Second, the optimal solutions to those optimization problems usually require dynamic programming with exponential computational complexity, which makes them impossible in practice. In this section, we propose using the utility

functions with respect to average waiting times in designing joint channel- and queue-aware scheduling, which was first reported in [68].

3.3.1 Utility Functions

Assume that user i is associated with an average waiting time W_i and the corresponding utility is $U_i(W_i)$. Obviously, with a long delay, the user has a low level of satisfaction (utility). It is reasonable to assume that $U_i(W_i)$ is decreasing. There are usually two approaches to obtaining utility functions. For a specific type of application, the utility function may be obtained by sophisticated subjective surveys. Another method is to design utility functions based on the habits of the traffic and appropriate fairness in the network.

3.3.2 Optimization Objective

Assume the average arrival bit rate of user i as λ_i, defined as

$$\lambda_i = \frac{1}{T_s} \lim_{n\to\infty} \frac{A_i[n]}{n}$$

where $A_i[n]$ is the total number of bits arriving during $(0, nT_s]$. Assuming that $Q_i[n]$ is ergodic, with Little's law, the average waiting time for user i, W_i, is

$$W_i = \frac{Q_i}{\lambda_i}$$

where $Q_i = \lim_{N\to\infty} \frac{\sum_{n=0}^{N-1} Q_i[n]}{N}$.

Let the base station control service bit rates so that

$$r_i[n]T_s \le Q_i[n]. \tag{3.7}$$

Then, the queue evolution equation (1.3) becomes

$$Q_i[n+1] = Q_i[n] - r_i[n]T_s + a_i[n]. \tag{3.8}$$

By exploiting an exponentially weighted low-pass filter, the average queue length, $\bar{Q}_i[n]$, can be updated as

$$\bar{Q}_i[n] = (1-\rho_w)\bar{Q}_i[n-1] + \rho_w Q_i[n] \tag{3.9}$$

where $0 < \rho_w < 1$.

Define the average waiting time over the time window at time nT_s as

$$W_i[n] = \frac{\bar{Q}_i[n]}{\lambda_i}. \tag{3.10}$$

At time nT_s (the beginning of time slot n), given the service rate $r_i[n]$, the predicted average waiting time at the end of time slot n, $(n+1)T_s$, is obtained by

$$\hat{W}_i[n+1] = \frac{\mathbb{E}_{a_i[n]}\{\bar{Q}_i[n+1]\}}{\lambda_i}$$

where $\mathbb{E}_{a_i[n]}\{\cdot\}$ denotes expectation with respect to $a_i[n]$. According to (1.3) and (3.9), we have

$$\mathbb{E}_{a_i[n]}\{\bar{Q}_i[n+1]\} = (1-\rho_w)\bar{Q}_i[n] + \rho_w(Q_i[n] - r_i[n]T_s + \mathbb{E}\{a_i[n]\})$$

Using $\mathbb{E}\{a_i[n]\} = \lambda_i T_s$, $\hat{W}_i[n+1]$, is obtained by

$$\begin{aligned}\hat{W}_i[n+1] &= (1-\rho_w)\frac{\bar{Q}_i[n]}{\lambda_i} + \rho_w\frac{Q_i[n]}{\lambda_i} + \rho_w T_s - \frac{\rho_w}{\lambda_i}T_s r_i[n] \\ &= (1-\rho_w)W_i[n] + \rho_w\frac{Q_i[n]}{\lambda_i} + \rho_w T_s - \frac{\rho_w}{\lambda_i}T_s r_i[n]\end{aligned}$$

Therefore, the predicted average waiting time at time $(n+1)T_s$ is a function of the service rate during time slot n, $r_i[n]$.

The optimization objective is to maximize the total utility with respect to the predicted average waiting times at each time slot in the network, that is,

$$\max_{r_i[n], i\in\mathcal{M}} \sum_{i=1}^{M} U_i(\hat{W}_i[n+1]).$$

Given the arrival processes, the average waiting time is actually determined by the service rate. It is obvious that

$$\frac{\partial U_i}{\partial r_i} = -\frac{\partial U_i}{\partial W_i}\frac{\rho_w}{\lambda_i}T_s.$$

If ρ_w is small enough, and using the properties of $U_i(W_i)$, we have

$$\begin{aligned}
&\sum_{i=1}^{M} U_i(\hat{W}_i[n+1]) - \sum_{i=1}^{M} U_i(\hat{W}_i[n]) \\
\approx &\sum_{i=1}^{M} \frac{\partial U_i}{\partial r_i}\Big|_{r_i=r_i[n-1]} (r_i[n] - r_i[n-1]) \\
\approx &\sum_{i=1}^{M} -\frac{\partial U_i}{\partial W_i}\Big|_{W_i=W_i[n]} \rho_w T_s(\frac{r_i[n]}{\lambda_i} - \frac{r_i[n-1]}{\lambda_i}) \\
= &\sum_{i=1}^{M} |\frac{\partial U_i}{\partial W_i}|\Big|_{W_i=W_i[n]} \rho_w T_s(\frac{r_i[n]}{\lambda_i} - \frac{r_i[n-1]}{\lambda_i})
\end{aligned}$$

Since the $r_i[n-1]$'s are fixed at time slot n, the optimization objective turns out to be a linear function of $r_i[n]$,

$$\max \sum_{i=1}^{M} \frac{\left|U_i'(W_i[n])\right|}{\lambda_i} r_i[n], \tag{3.11}$$

where $U_i'(W_i[n]) = \frac{\partial U_i(W_i)}{\partial W_i}\Big|_{W_i=W_i[n]}$, and $W_i[n]$ can be obtained from (3.10).

3.3.3 Problem Formulation in OFDM

If the subcarriers can dynamically be assigned, with the objective (3.11), we formulate this problem in the OFDM system as

$$\max_{D_i^{(n)}, i\in\mathcal{A}^n} \sum_{i\in\mathcal{A}^n} \frac{\left|U_i'(W_i[n])\right|}{\hat{\lambda}_i} r_i[n] \tag{3.12}$$

$$\text{subject to} \bigcup_{i\in\mathcal{A}^n} D_i^{(n)} \subseteq \mathcal{K}, \tag{3.13}$$

$$D_i^{(n)} \bigcap D_j^{(n)} = \varnothing, \quad i \neq j \quad \forall i,j \in \mathcal{A}^n, \tag{3.14}$$

$$r_i[n] \leq \frac{Q_i[n]}{T_s}, \quad i \in \mathcal{A}^n. \tag{3.15}$$

where the constraint (3.15) comes from the queue control rule (3.7), which means that the scheduler does not waste service rate. We refer to (3.15) as the *frugality constraint* (FC). Note that in the optimization objective (3.12) the estimated arrival rate $\hat{\lambda}_i$ replaces the expected (exact) value λ_i. This is because the base station does not know the arrival rates and the λ_i's must be estimated. They can also be estimated through an exponentially weighted low-pass window. Besides, there is another way to estimate the arrival rates.

Since the FC is applied, the scheduler does not serve any empty queue and waste service rate; therefore, λ_i equals the long-term average of the service rate of user i, $\mathbb{E}\{r_i\}$, in this scenario. In practice, we let

$$\hat{\lambda}_i = \bar{r}_i[n].$$

Letting

$$h(r; r_{max}) = \begin{cases} r & \text{if } r < r_{max}, \\ r_{max} & \text{if } r \geq r_{max}, \end{cases}$$

we can rewrite the optimization problem defined in (3.12)-(3.15) as

$$\max_{D_i^{(n)}, i \in \mathcal{A}^n} \sum_{i \in \mathcal{A}^n} \frac{\left|U_i'(\frac{\bar{Q}_i[n]}{\hat{\lambda}_i})\right|}{\hat{\lambda}_i} h(r_i[n]; \frac{Q_i[n]}{T_s}) \tag{3.16}$$

$$\text{subject to } \bigcup_{i \in \mathcal{A}^n} D_i^{(n)} \subseteq \mathcal{K}, \tag{3.17}$$

$$D_i^{(n)} \bigcap D_j^{(n)} = \varnothing \quad i \neq j \ \forall i, j \in \mathcal{A}^n. \tag{3.18}$$

3.3.4 Algorithms

The integer optimization problem (3.16)-(3.18) is NP-hard. In Section 2.3.1, an efficient and fast suboptimal DSA algorithm, a sorting-search algorithm, is proposed for the subcarrier assignment problem with the concave objective function. Note that the function $h(r; r_{max})$ is concave with respect to r. Therefore, the sorting-search algorithm can work well to solve the problem described in (3.16)-(3.18).

The FC is not necessary for the MDU scheduling. Without the FC, the MDU scheme can be implemented according to (3.4). To avoid ambiguity, we use MDU-FC to indicate the MDU working with the FC in this chapter. On the other hand, the FC can be applied in other scheduling schemes, such as those schemes mentioned in Section 3.2. Certainly, the sorting-search algorithm is needed in the case in which the FC is used.

3.4 Stability

It is shown in Chapter 2 that a utility function with respect to the data rate are directly associated with a kind of fairness. The trade-off between the spectral efficiency and fairness

is a core problem of resource allocation, especially for best-effort traffic. In addition, Chapter 2 demonstrates that concave utility functions can provide clear, tractable efficiency-fairness relations. Fortunately, a logarithmic function, which is concave, is usually used to describe best-effort traffic [30, 65].

Unlike best-effort traffic, the necessary condition for guaranteeing the QoS requirements of a delay-sensitive stream is that the service rate must be larger than the incoming rate of the stream. Therefore, the study of stability issue is the key to analyze scheduling algorithms for delay-sensitive traffic. In this chapter, we show the relationship between utility functions and stability. In fact, utility functions with very loose conditions (e.g. convexity/concavity is not required.) are able to stabilize the system using the MDU scheduling.

3.4.1 Background and Definition of Stability

The interaction between queueing and time-varying wireless channels is not well understood in a multiuser environment since multiple interacting queues result in difficulty in analysis. Currently, the stability property of scheduling is becoming more and more important [20, 49, 71]. First, the stability issue is essential for QoS provisioning and admission control. Moreover, the stability issue is mathematically tractable in many cases. There are two important methods to deal with the stability issue: Foster-Lyapunov drift [46] and fluid limit [17]. The Foster-Lyapunov mothed is classical for stability and harmonic analysis, but it may be very intricate in complicated scenarios. The fluid limit technique establishes the equivalency on stability between the orignal network and the associated fluid model with deterministic and continuous arrival streams. However, the above equivalent relationship for stability is usually built on the Markovian property of the system, and it is still unknown if the Markov assumption can be relaxed to just a stationary condition for the fluid limit technique in a general case. Moreover, both methods applied in most previous work such as [20, 45, 49, 71] (using the Foster-Lyapunov mothed) and [8, 63] (using the fluid limit technique) are challenged by the fact that the weights used in the MDU scheduling are functions of the current and previous queue states. In this section, we incorporate the concept and the properties of limit into the Foster-Lyapunov method to deal with the

above difficulty and to make the proofs concise.

For a queueing system, the system is stable if each queue length reaches a steady state and does not go to infinity. Mathematically, we define stability as follows. The system is stable if there exists $p > 0$ such that

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left\{\left|(\mathbf{Q}[n])^p\right|\right\} < \infty, \tag{3.19}$$

where $\mathbf{Q}[n] = (Q_1[n], Q_2[n], \cdots, Q_M[n])^T$, and for a vector $\mathbf{x} = [x_1, x_2, \cdots, x_M]^T$, $|\mathbf{x}| = \sum_{i=1}^{M} x_i$. To investigate the stability issue, we will first discuss the capacity region of the downlink system.

3.4.2 Capacity Region

Define a data rate vector $\mathbf{r}$ as

$$\mathbf{r} = (r_1, r_2, \ldots, r_M)^T \in \mathbb{R}_+^M,$$

where M is the number of users. The instantaneous capacity region for service data rates, $\mathcal{C}(\mathbf{H})$, is a set that consists of the total achievable data rate vectors in the current channel state $\mathbf{H}$. For instance, if DSA is allowed in the system, then the instantaneous data rate region is given by

$$\mathcal{C}_{DSA}(\mathbf{H}) = \left\{ \mathbf{r}(\mathbf{D}) : D_i \bigcap D_j = \varnothing,\ \forall i \neq j,\ \bigcup_{i\in\mathcal{M}} D_i \subseteq \mathcal{K} \right\},$$

where $\mathbf{D} = \{D_1, D_2, \cdots, D_M\}$. Usually, in practical systems, $\mathcal{C}(\mathbf{H})$ is a non-convex set since real systems can only provide finite modulation and coding schemes.

A resource allocation policy $\mathcal{R}(\mathbf{H})$ is said to be channel-stationary if the rate allocation depends only on the channel state $\mathbf{H}$. Note that channel-stationary policies can exploit time-sharing for the achievable data rate vectors in $\mathcal{C}(\mathbf{H})$. Hence, all available channel-stationary resource allocation policies can construct the convex hull of $\mathcal{C}(\mathbf{H})$; that is,

$$cov(\mathcal{C}(\mathbf{H})) = \{\mathcal{R}(\mathbf{H}) :\ \text{for all } \mathcal{R}\}. \tag{3.20}$$

Hence $cov(\mathcal{C}(\mathbf{H})$ can be seen as the capacity region that can be achieved by time-averaging two or more feasible rate vectors in the instantaneous capacity region in the channel state $\mathbf{H}$.

Assume the channel state process $\mathbf{H}(t)$ to be ergodic. Let $\tilde{\mathcal{C}}$ be the ergodic capacity region under the allocation constraints, which consists of the average data rate vectors obtained by all possible channel-stationary resource allocation schemes. Thus,

$$\tilde{\mathcal{C}} = \{\mathbb{E}\{\mathcal{R}(\mathbf{H})\} : \text{ for all } \mathcal{R}\}.$$

Explicitly, the ergodic capacity region is a closed, convex, and compact set. Nevertheless, we do not consider non-channel-stationary resource allocation policies, for which the average service data rate vector is defined as

$$\liminf_{t\to\infty} \frac{\int_{\tau=0}^{t} \mathbf{r}(\tau)\, d\tau}{t}.$$

However, we have the following lemma.

Lemma 3.1 *With ergodic channel state processes, any average service data rate vector under any non-channel-stationary policy still lies in the ergodic capacity region* $\tilde{\mathcal{C}}$.

The proof is shown in Appendix D.

The lemma claims that the long-term average service rate vector under any resource allocation policy lies in the ergodic capacity region, which is determined by the physical layer techniques and the channel distributions.

3.4.3 Maximum Stability Region

Assume that the input streams are stationary and ergodic with rate vector $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_M]^T$, and that the channel processes are stationary and ergodic as well. Then, with a similar proof to that of Lemma 1b in [49], we have the following lemma.

Lemma 3.2 *The necessary condition for stability is* $\boldsymbol{\lambda} \in \tilde{\mathcal{C}}$.

However, in the case $\boldsymbol{\lambda} \in \tilde{\mathcal{C}}$, not all scheduling policies can stabilize the system. The *stability region* of a policy is defined to be the set of all possible arrival rate vectors for which the system is stable under the policy [71]. Note that the capacity region is concerned with the service data rates, whereas the stability region is with regard to the arrival rates. The *maximum stability region* is defined as the largest stability region that can be achieved

by some scheduling schemes. Similarly, a policy is called a maximum-stability-region policy if the stability region of the policy covers all stability regions under all other policies. Thus, the concept of maximum-stability-region policy is interchangeable with the concept of throughput-optimal policy in [8].

Naturally, we are interested in the following questions. First, does the maximum stability region always exist? Second, how large can the maximum stability region be? Finally, how do we identify and design maximum-stability-region policies without the statistical information about the arrivals and the wireless channels? The answer to the first question is yes. Mathematically, the maximum stability region is the superset of stability regions of all possible policies. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the stability regions of policy $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. Then, we can construct a policy $\mathcal{R}$ such that

$$\mathcal{R} = \begin{cases} \mathcal{R}_1 & \text{if } \boldsymbol{\lambda} \in \mathcal{S}_1 - \mathcal{S}_2 \\ \mathcal{R}_2 & \text{if } \boldsymbol{\lambda} \in \mathcal{S}_2 - \mathcal{S}_1 \\ \mathcal{R}_1 \text{ or } \mathcal{R}_2 & \text{if } \boldsymbol{\lambda} \in \mathcal{S}_1 \bigcap \mathcal{S}_2. \end{cases}$$

Thus, the policy $\mathcal{R}$ has the stability region $\mathcal{S}_1 \bigcup \mathcal{S}_2$. With the same method, we can always construct a policy with the superset of stability regions of all possible policies - the maximum stability region. For the second question, it is easy to show from Lemma 3.2 that the maximum stability region must be a subset of $\tilde{\mathcal{C}}$. We will explore the remaining questions by investigating a more general scheduling policy that allocates data rate vectors such that

$$\max_{\mathbf{r}[n] \in \mathcal{C}(\mathbf{H}[n])} \underline{g}^T \left(\mathbf{V}[n]\right) \mathbf{r}[n] \tag{3.21}$$

where the vector function $\underline{g}(\cdot)$ and the vector $\mathbf{V}[n]$ are described as follows:

- Let $\underline{g}(\mathbf{x}) = [g_1(x_1), g_2(x_2), \cdots, g_M(x_M)]^T$ and assume that the functions $g_i(\cdot)$'s are non-negative and non-decreasing functions such that

$$\text{for } x < \infty,\ g_i(x) < \infty, \tag{3.22}$$

$$\lim_{x \to \infty} g_i(x) = \infty, \tag{3.23}$$

and given any constant $A > 0$,

$$\limsup_{x\to\infty} \frac{g_i(x+A)}{g_i(x)} = 1. \tag{3.24}$$

- Let $\mathbf{V}[n] = (V_1[n], V_2[n], \cdots, V_M[n])^T$, where $V_i[n] = f(Q_i[n], Q_i[n-1], \cdots)$. The function f is non-negative and non-decreasing with respect to the $Q_i[n]$'s for all n. Furthermore,

$$\mathbb{E}\{|Q_i[n] - V_i[n]|\} < \infty \text{ for all } i. \tag{3.25}$$

- In addition to the ergodicity of the channel processes and arrival streams, we assume that

$$\mathbb{E}\left\{g_i(a_i[n])a_i[n]\right\} < \infty \text{ for all } i. \tag{3.26}$$

  where $a_i[n]$ is the arrival bits during a time slot for user i. From a practical point of view, any achievable instantaneous data rates are bounded, which can also simplify the proof of the following theorem.

We first consider the optimization problem given by

$$\max_{\mathbf{r}\in\mathcal{C}(\mathbf{H}[n])} \mathbf{w}^T\mathbf{r}, \tag{3.27}$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_M]^T$. Clearly, the scheduling policy based on (3.27) is stationary. Furthermore, the following lemma shows that the scheduling policy also leads to optimality in the long-term sense.

Lemma 3.3 *For a given weight vector* $\mathbf{w}$*, assume that* $\mathbf{r}^*(\mathbf{H})$ *is the optimization problem (3.27) in the instantaneous capacity region* $\mathcal{C}(\mathbf{H})$*. Let* $\tilde{\mathbf{r}}^* = \mathbb{E}\{\mathbf{r}^*(\mathbf{H})\}$*, then* $\tilde{\mathbf{r}}^*$ *is the optimal solution to the following optimization problem in the ergodic capacity region* $\tilde{\mathcal{C}}$

$$\max_{\tilde{\mathbf{r}}\in\tilde{\mathcal{C}}} \mathbf{w}^T\tilde{\mathbf{r}}. \tag{3.28}$$

The proof is shown in Appendix E. Then, we present the major results for the stability issue in the following theorem.

Theorem 3.1 *If the average arrival rate vector is within the interior of the ergodic capacity region, Int($\tilde{\mathcal{C}}$), where Int($\tilde{\mathcal{C}}$) = $\tilde{\mathcal{C}}$ − the boundary of $\tilde{\mathcal{C}}$, then the scheduling (3.21) satisfying the conditions (3.22) - (3.26) stabilizes the queues in the following sense*

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left\{\left|\underline{g}(\mathbf{V}[n])\right|\right\} < \infty. \tag{3.29}$$

In other words, the scheduling has the maximum stability region, which is Int($\tilde{\mathcal{C}}$).

The proof is shown in Section 3.5.

Note that the performance of the scheduling (3.21) is worse than that of the scheduling (3.21) with the FC since the scheduling (3.21) may waste some subcarriers on empty queues. The system without the use of the FC is called the *dominant system*. Therefore, the scheduling (3.21) with the FC has the maximum stability region as well.

To study the MDU scheduling, we have to prove the relation (3.25) first. We have the following lemma.

Lemma 3.4 *Let $\bar{Q}_i[0] = Q_i[0]$ for all i. Then*

$$\mathbb{E}\{|Q_i[n] - \bar{Q}_i[n]|\} < \infty \text{ for all } i.$$

The proof is shown in Appendix F. The following corollary states the stability property of the MDU scheduling.

Corollary 3.1 *Express the weights $\frac{|U_i'(\bar{Q}_i[n]/\hat{\lambda}_i)|}{\hat{\lambda}_i}$'s in the MDU scheduling as the $g_i(\bar{Q}_i[n])$'s. Then the MDU scheduling with the functions $g_i(\cdot)$'s satisfying the conditions (3.22) - (3.26) has the maximum stability region, Int($\tilde{\mathcal{C}}$). If the average arrival rate vector are within Int($\tilde{\mathcal{C}}$), then the MDU scheduling policy is stable, that is,*

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left\{\left|\underline{g}(\rho_w \mathbf{Q}[n])\right|\right\} < \infty,$$

where $\rho_w \mathbf{Q}[n] = (\rho_w Q_1[n], \rho_w Q_2[n], \cdots, \rho_w Q_M[n])^T$.

Proof: The weights of the MDU scheduling are $g_i(V_i[n])$'s, where $V_i[n] = \bar{Q}_i[n]$. Lemma 3.4 shows the validity of (3.25) for the MDU scheduling. It follows from (3.9) that

$$\rho_w \mathbf{Q}[n] \le \bar{\mathbf{Q}}[n].$$

Since $\underline{g}(\cdot)$ is non-decreasing, then

$$\underline{g}(\rho_w \mathbf{Q}[n]) \leq \underline{g}(\bar{\mathbf{Q}}[n]).$$

Therefore, we obtain

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left\{\left|\underline{g}(\rho_w \mathbf{Q}[n])\right|\right\} &\leq \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}\left\{\left|\underline{g}(\bar{\mathbf{Q}}[n])\right|\right\} \\ &< \infty.\end{aligned}$$

□

Remarks

- The general scheduling rule (3.21) that is able to achieve the maximum stability region does not require statistical information about the arrivals and the wireless channels.

- If $g_i(x)$ is continuously differentiable, the condition (3.24) can be replaced by

  $$\lim_{x\to\infty} \frac{g_i'(x)}{g_i(x)} = 0.$$

  Intuitively, any non-negative and increasing function whose increasing order is higher than or equal to the logarithm function and lower than the exponential function satisfies both conditions (3.23) and (3.24). Therefore, there are many degrees of freedom for designing scheduling policies with the maximum stability region. Similarly, there is enough room to choose the $V_i[n]$'s. For instance, given a finite positive integer J, $V_i[n] = Q_i[n-J]$, $V_i[n] = \sum_{j=0}^{J-1} Q_i[n-j]/J$, $V_i[n] = (Q_i[n] \cdot Q_i[n-2] \cdot Q_i[n-J+1])^{\frac{1}{J}}$, etc., are all able to stabilize the system.

- The maximum stability region is shown to be the interior of the ergodic capacity region that is determined by the physical layer techniques. Figure 3.1 illustrates the stability regions of some scheduling schemes for the two-user case. According to Lemma 3.3, the rate allocation of the MSC scheduling is the solution to the following problem

  $$\max_{[\tilde{r}_1,\tilde{r}_2]^T \in \tilde{\mathcal{C}}} \tilde{r}_1 + \tilde{r}_2.$$

Thus, the optimal solution $[\tilde{r}_1^*, \tilde{r}_2^*]^T$ should be the tangent point between the boundary of $\tilde{\mathcal{C}}$ and a line $\tilde{r}_1+\tilde{r}_2 = b$ for an appropriate b; the stability region is, therefore, $\lambda_1 < \tilde{r}_1^*$ and $\lambda_2 < \tilde{r}_2^*$. The PF scheduling with a very small ρ_w leads to the optimization problem

$$\max_{[\tilde{r}_1,\tilde{r}_2]^T \in \tilde{\mathcal{C}}} \frac{1}{\tilde{r}_1^\dagger}\tilde{r}_1 + \frac{1}{\tilde{r}_2^\dagger}\tilde{r}_2.$$

Similar to the MSC, the optimal rate vector for the PF $[\tilde{r}_1^\dagger, \tilde{r}_2^\dagger]^T$ should be the point of tangency between the boundary of $\tilde{\mathcal{C}}$ and a line $\frac{1}{\tilde{r}_1^\dagger}\tilde{r}_1 + \frac{1}{\tilde{r}_2^\dagger}\tilde{r}_2 = b'$ with an appropriate b'. Since the MSC and PF scheduling schemes have small stability regions, they cannot stabilize all arrival vectors inside the ergodic capacity region but outside their stability regions.

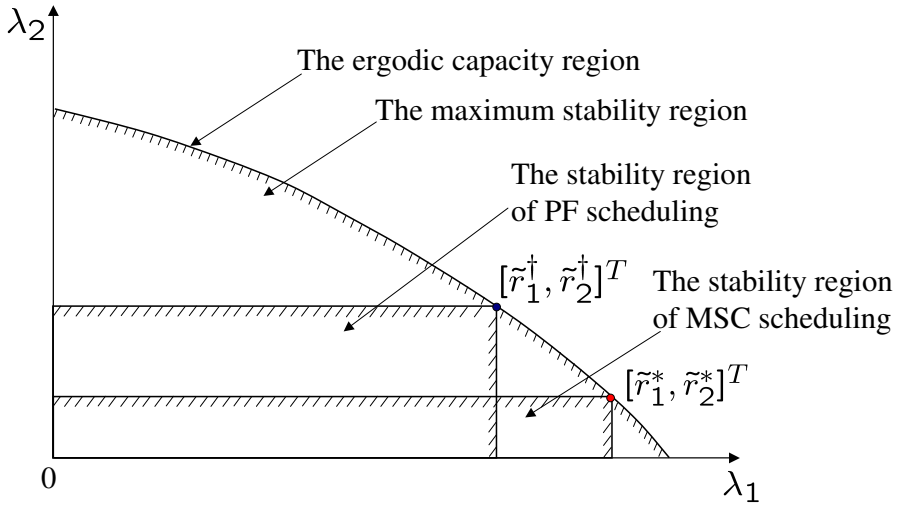


Figure 3.1. Stability regions for different scheduling schemes in the two-user case

- In the proof of Theorem 3.1, we see that the FC cannot stabilize scheduling approaches without the maximum stability region, and that the effect of the FC may become marginal with a heavy traffic load. The impact of the FC on different scheduling policies will be discussed in Section 3.7.

- To obtain more system gains, we should jointly design and optimize techniques in multiple layers, but cross-layer design usually seems complicated and not transparent. However, the above result gives us a guideline for cross-layer optimization. First, use

advanced physical layer techniques to enlarge the ergodic capacity region. Second, design a scheduling scheme with the maximum stability region to fully exploit the ergodic capacity region.

3.5 Proof of Theorem 3.1

The primary method used in the proof is the Foster-Lyapunov method. A new tip is to apply Fatou's lemma and the definition of the upper limit. The proof does not require the Markovian property on the channel states and/or the arrival traffic.

Let the Lyapunov function be

$$L(\mathbf{Q}[n]) = \sum_{i \in \mathcal{M}} L_i(Q_i[n]),$$

where $\frac{dL_i(x)}{dx} = g_i(x)$. Define

$$\begin{aligned} \mathbf{Q}'[n+1] &= \mathbf{Q}[n] - \mathbf{r}[n]T_s + \mathbf{a}[n] \\ \text{and } \boldsymbol{\xi} &= -\mathbf{r}[n]T_s + \mathbf{a}[n]. \end{aligned} \tag{3.30}$$

Using the mean value theorem [57], we obtain

$$\begin{aligned} L(\mathbf{Q}'[n+1]) - L(\mathbf{Q}[n]) &= \nabla L^T(\mathbf{Q}'[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi})\boldsymbol{\xi} \\ &= \underline{g}^T(\mathbf{Q}'[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi})\boldsymbol{\xi}, \end{aligned}$$

where $\mathbf{a} \odot \mathbf{b} = [a_1 b_1, a_2 b_2, \cdots, a_M b_M]^T$, $0 < \nu_i < 1$ for all i. Clearly,

$$\begin{aligned} & L(\mathbf{Q}'[n+1]) - L(\mathbf{Q}[n]) \\ &= \underline{g}^T(\mathbf{V}[n])\boldsymbol{\xi} + \left[\underline{g}(\mathbf{Q}[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi}) - \underline{g}(\mathbf{V}[n])\right]^T \boldsymbol{\xi}. \end{aligned} \tag{3.31}$$

Define the state pair $\mathbf{Y}[n] = (\mathbf{Q}[n], \mathbf{V}[n])$. Conditioning on $\mathbf{Y}[n]$ and taking expectation, we obtain

$$\begin{aligned} & \mathbb{E}\left\{L(\mathbf{Q}'[n+1]) - L(\mathbf{Q}[n]) \middle| \mathbf{Y}[n]\right\} \\ &= \underline{g}^T(\mathbf{V}[n])\mathbb{E}\left\{\boldsymbol{\xi} \middle| \mathbf{Y}[n]\right\} && (3.32) \\ &\quad + \mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi}) - \underline{g}(\mathbf{V}[n])\right]^T \boldsymbol{\xi} \middle| \mathbf{Y}[n]\right\}. && (3.33) \end{aligned}$$

We will study parts (3.32) and (3.33) separately. (3.32) then becomes

$$\begin{aligned}\underline{g}^T(\mathbf{V}[n])\mathbb{E}\left\{\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\} &= T_s\underline{g}^T(\mathbf{V}[n])\mathbb{E}\left\{\frac{\mathbf{a}[n]}{T_s}-\mathbf{r}[n]\middle|\mathbf{Y}[n]\right\}\\ &= T_s\underline{g}^T(\mathbf{V}[n])\left(\boldsymbol{\lambda}-\mathbb{E}\left\{\mathbf{r}[n]\middle|\mathbf{Y}[n]\right\}\right). \end{aligned} \tag{3.34}$$

According to Lemma 3.3, the scheduling policy (3.21) results in

$$\mathbb{E}\left\{\mathbf{r}[n]\middle|\mathbf{Y}[n]\right\} = \arg\max_{\tilde{\mathbf{r}}[n]\in\tilde{\mathcal{C}}}\underline{g}^T(\mathbf{V}[n])\tilde{\mathbf{r}}[n],$$

which minimizes (3.34). Since $\boldsymbol{\lambda}$ is located within the interior ergodic capacity region $\tilde{\mathcal{C}}$, there exists a rate vector $\mathbf{r}'\in\tilde{\mathcal{C}}$ such that $r_i'>\lambda_i$ for all i. Let $\delta=\min_i(r_i'-\lambda_i)$. Thus, under the scheduling policy,

$$\begin{aligned}\underline{g}^T(\mathbf{V}[n])\mathbb{E}\left\{\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\} &\le T_s\underline{g}^T(\mathbf{V}[n])\left(\boldsymbol{\lambda}-\mathbf{r}'[n]\right)\\ &< -T_s|\underline{g}(\mathbf{V}[n])|\delta. \end{aligned} \tag{3.35}$$

To explore the property of (3.33), we consider

$$\limsup_{\mathbf{V}[n]\to\infty}\frac{\mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\}}{|\underline{g}(\mathbf{V}[n])|}. \tag{3.36}$$

Using the dual of Fatou's lemma [26], we obtain

$$\begin{aligned}(3.36) &= \limsup_{\mathbf{V}[n]\to\infty}\mathbb{E}\left\{\frac{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}}{|\underline{g}(\mathbf{V}[n])|}\right\}\\ &\le \mathbb{E}\left\{\limsup_{\mathbf{V}[n]\to\infty}\frac{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}}{|\underline{g}(\mathbf{V}[n])|}\right\}\\ &\le \mathbb{E}\left\{\limsup_{\mathbf{V}[n]\to\infty}\sum_{i\in\mathcal{M}}\frac{\left[g_i(Q_i[n]+\nu_i\xi_i)-g_i(V_i[n])\right]\xi_i}{g_i(V_i[n])}\right\}.\\ &= \mathbb{E}\left\{\sum_{i\in\mathcal{M}}\left[\limsup_{V_i[n]\to\infty}\frac{g_i(Q_i[n]+\nu_i\xi_i)}{g_i(V_i[n])}-1\right]\xi_i\right\}\end{aligned}$$

Let $\boldsymbol{\zeta}=\mathbf{Q}[n]-\mathbf{V}[n]$, then $\mathbb{E}\{|\zeta_i|\}<\infty$ for all i according to the condition (3.25). Equation (3.30) and the condition (3.26) lead to

$$\begin{aligned}\mathbb{E}\{|\xi_i|\} &\le \mathbb{E}\{r_i[n]\}T_s+\mathbb{E}\{a_i[n]\}\\ &< \infty \text{ for all } i.\end{aligned}$$

Hence it follows from the properties of the $g_i(\cdot)$'s given by (3.23) and (3.24) that

$$\begin{aligned}
&\limsup_{V_i[n]\to\infty} \frac{g_i(Q_i[n]+\nu_i\xi_i)}{g_i(V_i[n])} \\
&= \limsup_{V_i[n]\to\infty} \frac{g_i(V_i[n]+\zeta_i+\nu_i\xi_i)}{g_i(V_i[n])} \\
&= 1 \text{ with probability 1.}
\end{aligned}$$

Therefore, we have

$$\left[\limsup_{V_i[n]\to\infty} \frac{g_i(Q_i[n]+\nu_i\xi_i)}{g_i(V_i[n])} - 1\right]\xi_i = 0 \text{ with probability 1,}$$

and

$$\begin{aligned}
&\limsup_{\mathbf{V}[n]\to\infty} \frac{\mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\}}{|\underline{g}(\mathbf{V}[n])|} \\
&\le \mathbb{E}\left\{\limsup_{\mathbf{V}[n]\to\infty}\sum_{i\in\mathcal{M}}\frac{[g_i(Q_i[n]+\nu_i\xi_i)-g_i(V_i[n])]\,\xi_i}{g_i(V_i[n])}\right\} \\
&= 0,
\end{aligned}$$

which means that

$$\limsup_{\mathbf{V}[n]\to\infty} \frac{\mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\}}{|\underline{g}(\mathbf{V}[n])|} = -\Omega_0,$$

where $\Omega_0 \ge 0$.

The definition of the upper limit [2] implies that for any $\epsilon > 0$, there exists $\mathbf{V}^* > 0$ such that for $\mathbf{V}[n] > \mathbf{V}^*$,

$$\frac{\mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\}}{|\underline{g}(\mathbf{V}[n])|} < -\Omega_0+\epsilon \tag{3.37}$$

$$< \epsilon. \tag{3.38}$$

It follows from (3.37) and the fact that $|\underline{g}(\mathbf{V}[n])| > 0$ for $\mathbf{V}[n] > \mathbf{V}^*$ that

$$\mathbb{E}\left\{\left[\underline{g}(\mathbf{Q}[n]+\boldsymbol{\nu}\odot\boldsymbol{\xi})-\underline{g}(\mathbf{V}[n])\right]^T\boldsymbol{\xi}\middle|\mathbf{Y}[n]\right\} < \epsilon|\underline{g}(\mathbf{V}[n])|. \tag{3.39}$$

[2] $\limsup_{x\to\infty} f(x) = \lim_{y\to\infty}\sup\{f(x): x > y\} = \inf_y \sup\{f(x): x > y\}$

Due to the assumptions (3.22) and (3.26), there must exist a positive number $\Omega_1 < \infty$ such that

$$\sup_{\mathbf{V}[n] \le \mathbf{V}^*} \mathbb{E}\left\{ \left[\underline{g}(\mathbf{Q}[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi}) - \underline{g}(\mathbf{V}[n])\right]^T \boldsymbol{\xi} \middle| \mathbf{Y}[n] \right\} < \Omega_1. \tag{3.40}$$

Part (3.33) can be obtained from (3.39) and (3.40) as

$$\mathbb{E}\left\{ \left[\underline{g}(\mathbf{Q}[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi}) - \underline{g}(\mathbf{V}[n])\right]^T \boldsymbol{\xi} \middle| \mathbf{Y}[n] \right\} < \begin{cases} \Omega_1 & \mathbf{V}[n] \le \mathbf{V}^* \\ \epsilon |\underline{g}(\mathbf{V}[n])| & \mathbf{V}[n] > \mathbf{V}^* \end{cases},$$

which can be combined into

$$\mathbb{E}\left\{ \left[\underline{g}(\mathbf{Q}[n] + \boldsymbol{\nu} \odot \boldsymbol{\xi}) - \underline{g}(\mathbf{V}[n])\right]^T \boldsymbol{\xi} \middle| \mathbf{Y}[n] \right\} < \epsilon |\underline{g}(\mathbf{V}[n])| + \Omega_1. \tag{3.41}$$

Therefore, it follows from (3.35) and (3.41) that

$$\mathbb{E}\left\{ L(\mathbf{Q}'[n+1]) - L(\mathbf{Q}[n]) \middle| \mathbf{Y}[n] \right\} < -(T_s\delta - \epsilon)|\underline{g}(\mathbf{V}[n])| + \Omega_1.$$

Since ϵ is an arbitrary positive, we let ϵ be small enough so that $\epsilon < T_s\delta$.

On the other hand,

$$\mathbf{Q}[n+1] - \mathbf{Q}'[n+1] = \begin{cases} 0 & \mathbf{r}[n]T_s \le \mathbf{Q}[n] \\ \mathbf{r}[n]T_s - \mathbf{Q}[n] & \mathbf{r}[n]T_s > \mathbf{Q}[n], \end{cases}$$

and

$$\begin{aligned} &\mathbb{E}\left\{ L(\mathbf{Q}[n+1]) - L(\mathbf{Q}'[n+1]) \middle| \mathbf{Y}[n] \right\} \\ &= \mathbb{E}\left\{ \underline{g}\left(\mathbf{Q}'[n+1] + \boldsymbol{\nu}' \odot (\mathbf{r}[n]T_s - \mathbf{Q}[n])\right) (\mathbf{r}[n]T_s - \mathbf{Q}[n]) \cdot \mathbf{1}\{\mathbf{r}[n]T_s > \mathbf{Q}[n]\} \middle| \mathbf{Y}[n] \right\}, \end{aligned}$$

where $\mathbf{1}\{\text{event}\}$ equals 1 if the event is true, whereas it equals 0. Therefore, $\mathbb{E}\left\{ L(\mathbf{Q}[n+1]) - L(\mathbf{Q}'[n+1]) \middle| \mathbf{Y}[n] \right\}$ is bounded by a positive number, Ω_2.

Consequently, it follows that

$$\begin{aligned} &\mathbb{E}\left\{ L(\mathbf{Q}[n+1]) - L(\mathbf{Q}[n]) \middle| \mathbf{Y}[n] \right\} \\ &= \mathbb{E}\left\{ L(\mathbf{Q}[n+1]) - L(\mathbf{Q}'[n+1]) \middle| \mathbf{Y}[n] \right\} + \mathbb{E}\left\{ L(\mathbf{Q}'[n+1]) - L(\mathbf{Q}[n]) \middle| \mathbf{Y}[n] \right\} \\ &< -(T_s\delta - \epsilon)\left|\underline{g}(\mathbf{V}[n])\right| + \Omega_1 + \Omega_2. \end{aligned}$$

Taking expectation with respect to $\mathbf{Y}[n]$, we obtain

$$\mathbb{E}\left\{L(\mathbf{Q}[n+1]) - L(\mathbf{Q}[n])\right\} < -(T_s\delta - \epsilon)\mathbb{E}\left\{\left|\underline{g}(\mathbf{V}[n])\right|\right\} + \Omega_1 + \Omega_2.$$

Taking summation, we have

$$\mathbb{E}\left\{L(\mathbf{Q}[N])\right\} - L(\mathbf{Q}[0]) < -(T_s\delta - \epsilon)\sum_{n=0}^{N-1}\mathbb{E}\left\{\left|\underline{g}(\mathbf{V}[n])\right|\right\} + N(\Omega_1 + \Omega_2),$$

and

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\{\left|\underline{g}(\mathbf{V}[n])\right|\right\} < \frac{\Omega_1 + \Omega_2}{T_s\delta - \epsilon} + \frac{1}{N}L(\mathbf{Q}[0]).$$

Let $N \to \infty$, $\frac{1}{N}L(\mathbf{Q}[0]) \to 0$, and then the theorem is proved.

3.6 Further Improvement Through Delay Transmit Diversity and Adaptive Power Allocation

The studies on the stability issue in Section 3.4 can directly guide us to techniques for improving the performance of multicarrier scheduling. According to the properties of the maximum stability region, we propose the use of joint stabilizing scheduling and power allocation to extend the maximum stability region so as to enhance the throughput-delay performance. Moreover, we propose the use of delay transmit diversity to increase the fluctuations in the frequency domain, by which we can obtain more frequency diversity.

3.6.1 Joint Dynamic Subcarrier Assignment and Adaptive Power Allocation

In Section 3.4, we showed that the maximum stability region is the interior of the ergodic capacity region at the physical layer. Thus, any techniques that are able to enlarge the ergodic capacity region can definitely improve the system performance. Adaptive power allocation lets the transmit power at each subcarrier be adjustable and only constrained by the total power limit $\bar{P}$; let $p[k]$ be the power at subcarrier k, then $\sum_{k\in\mathcal{K}} p[k] \le \bar{P}$. Then, the instantaneous data rate region for the joint DSA and APA is given by

$$\mathcal{C}_{DSA+APA}(\mathbf{H}) = \left\{\mathbf{r}(\mathbf{D},\mathbf{p}) : D_i \bigcap D_j = \varnothing,\ \forall i \neq j,\ \bigcup_{i\in\mathcal{M}} D_i \subseteq \mathcal{K}, \sum_{k\in\mathcal{K}} p[k] \le \bar{P}\right\},$$

where $\mathbf{p} = \{p[1], p[2], \cdots, p[K]\}$. Since the joint DSA and APA has looser constraints on resource allocation than the DSA, $\mathcal{C}_{DSA}(\mathbf{H}) \subseteq \mathcal{C}_{DSA+APA}(\mathbf{H})$ for each channel state $\mathbf{H}$;

therefore, the ergodic capacity region of joint DSA and APA is larger than that of DSA as well. As long as a scheduling scheme with the maximum stability region is applied in the system, the enlarged ergodic capacity region can be fully exploited. Obviously, the scheduling rule (3.21) on $\mathcal{C}_{DSA+APA}(\mathbf{H})$ still has the maximum stability region.

The effect of APA is influenced by the rate adaptation used in the system. With continuous rate adaptation, the improvement of APA is trivial; however, the improvement of APA becomes substantial when there are only a small number of modulation levels, which is shown in Chapter 2.

For discrete rate adaptation, water-filling is not optimal for power allocation. In Section 2.3.2.3, a greedy power allocation algorithm is proposed to achieve the optimality of optimization problems with a concave objective function. To solve the joint DSA and APA problem, the sorting-search DSA and the greedy APA algorithms can be used iteratively. Mathematically, the MDU with FC can be expressed as the above optimization problem; thus, the sorting-search DSA and the greedy APA algorithms proposed in Section 2.3 can be implemented with no change.

3.6.2 Delay Transmit Diversity

In a single-carrier network, the multiuser diversity gain is limited in environments with little scattering or slow fading. Opportunistic beamforming is proposed in [76] to induce fast and large fluctuations so as to amplify the multiuser diversity gain. The main idea of opportunistic beamforming is to change the magnitudes and phases of antenna weights in a pseudorandom fashion.

In a multicarrier network, the multiuser diversity gain is diminished in environments with flat fading, which is usually caused by a line-of-sight path and/or little scattering. Therefore, we propose a simpler multiple transmit antenna scheme, *delay transmit diversity*, to increase the randomness in the frequency domain compared with the opportunistic beamforming in the time domain.

Delay transmit diversity was first proposed in single-carrier systems [78]. Delay transmit diversity actually converts spatial diversity into frequency diversity by inducing multiple

paths. Thus, the Viterbi algorithm is needed in the receivers in single-carrier systems. An OFDM system with delay transmit diversity is shown in Figure 3.2. The signals from

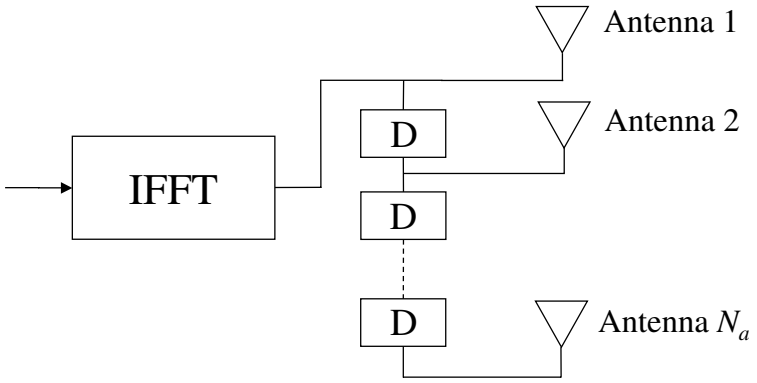


Figure 3.2. Delay transmit diversity in an OFDM system

additional antennas are the same as the signal from the first antenna but with different delays. Note that delay transmit diversity is implemented after *inverse fast Fourier transform* (IFFT) processing. For traditional OFDM systems, although the delay transmit diversity does not require additional processing in the receivers, some channel coding across subcarriers is needed to obtain the frequency diversity amplified by the delay transmit diversity [38]. However, in OFDM systems using DSA, the delay transmit diversity becomes totally transparent since any modulation, coding, and scheduling schemes used in the single-antenna case remain unchanged. Moreover, the frequency diversity induced by delay transmit diversity is transformed into multiuser diversity, which can be absorbed through DSA. In brief, delay transmit diversity and opportunistic beamforming are duals of each other.

3.7 Simulation Results and Performance Comparison

In this section, we compare the performance of different scheduling schemes in an OFDM network. In the simulation, each user's channel suffers multipath Rayleigh fading with the delay profile of Channel B for outdoor to indoor and pedestrian environments in [54], and each user is assumed to be stationary or slowly moving so that the maximum Doppler shift is 10 Hz. In the OFDM network, there are 128 subcarriers in a total channel bandwidth of 1.920 MHz. We assume that there are 20 users in the system. These 20 users have different

distances from the base station; consequently, their average achievable transmission rates are different due to path loss.

Let the acceptable BER be 10^{-6} for rate adaptation since data transmission is sensitive to error. Assume that a set of achievable transmission rates in bits/sec per Hz is {0,1/2,1,2,3,4}. The transmission rate is chosen to be the largest available rate whose required SNR determined by (2.2) is larger than or equal to the current SNR. In practice, we can use 1/2-rate channel coding and a series of modulation schemes including BPSK, QPSK, 16-QAM, 64-QAM, as well as 256-QAM to achieve the above feasible rates.

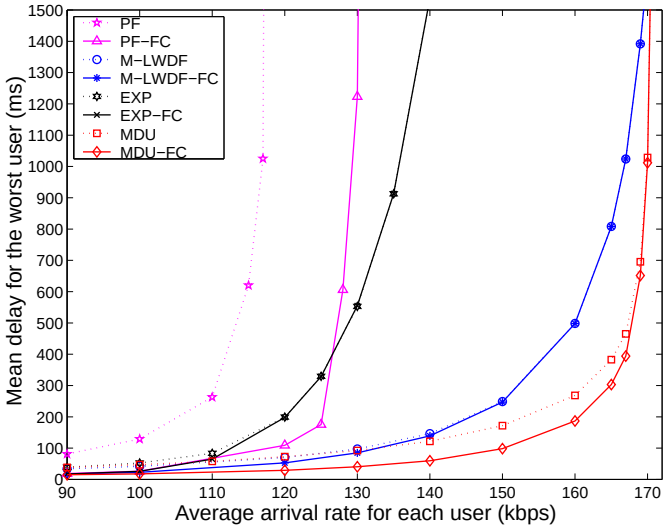
The packet length is assumed to be independently and exponentially distributed with an average length of 1024 bits. The packet arrival is modeled as a Poisson random variable for the following two reasons. First, delay-sensitive traffic is usually generated smoothly. Second, delays in the system are determined by two factors: the burst of arrivals and the fluctuation of scheduled service rates. Since in this chapter we are more interested in the second factor, Poisson arrival is assumed.

The length of a time slot T_s is 4 ms. All simulations were run for 300,000 slots, which correspond to 20 minutes in reality.

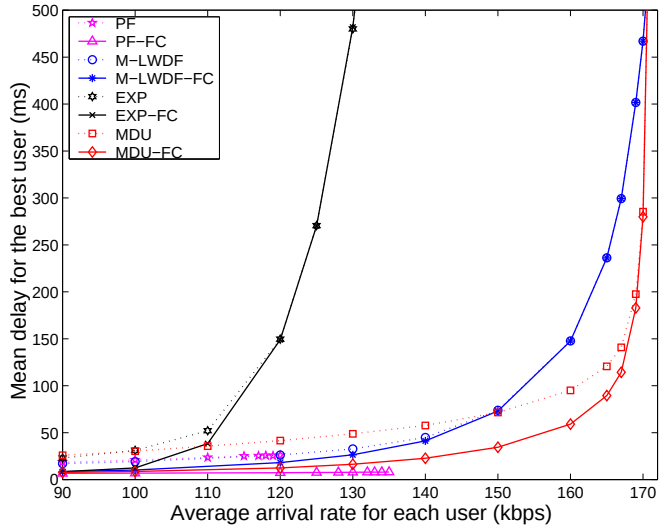
3.7.1 Performance Comparison

The simulation results are shown in Figure 3.3 in terms of traffic load versus mean delay. In each simulation, all users have the same arrival rate. Due to the asymmetry among users' channel conditions, Figure 3.3 represents the mean delays for the worst user who has the smallest average SNR, for the best user who has the largest average SNR, and averaging for all users in the system, respectively. We compare the performance of PF, M-LWDF, EXP, and MDU scheduling rules in an adaptive OFDM network. For comparison, the MDU uses $|U_i'(W)| = W$ for all i. Since the FC can be deployed with all scheduling rules, we run two schemes for each scheduling rule: with and without FC, respectively.

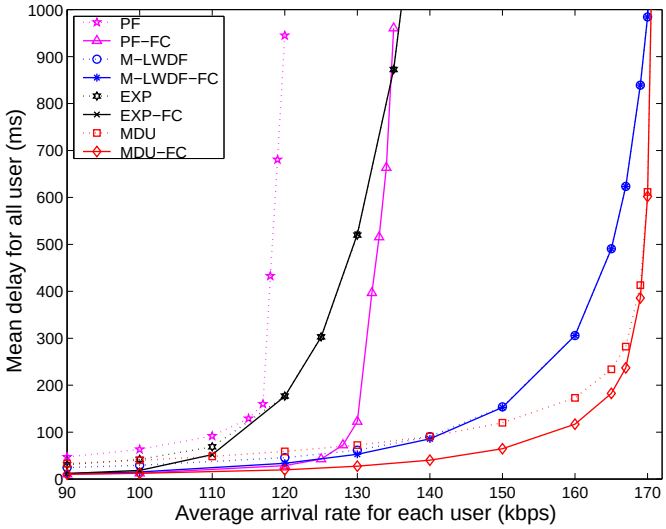
Figure 3.3 shows the advantages of maximum-stability-region scheduling schemes. Since the PF scheduling has a small stability region, the maximum throughput the PF is able to support for the worst user is around 117 kbps, whereas the saturated throughput of the



(a) The average waiting time for the worst user



(b) The average waiting time for the best user



(c) The average waiting time for all users

Figure 3.3. Delay performance of different scheduling policies

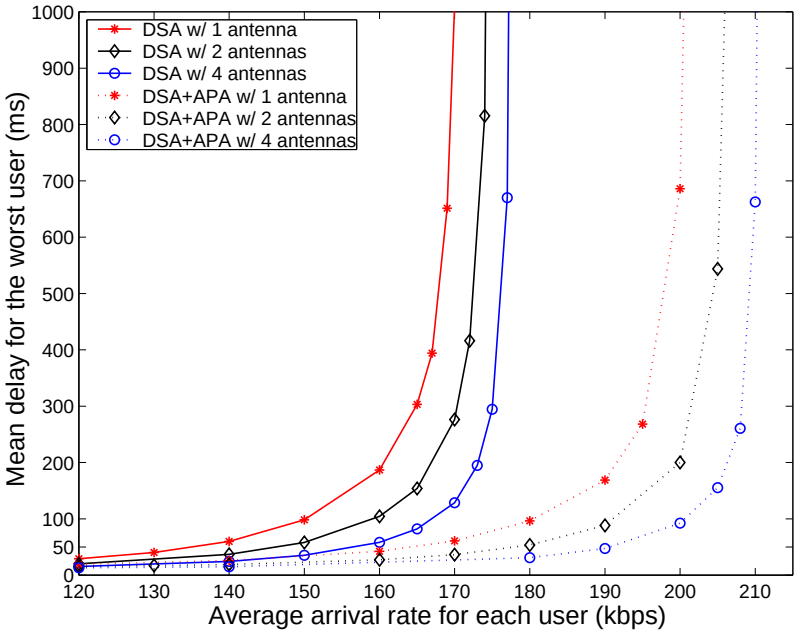
worst user with the MDU is 170 kbps. Furthermore, the PF scheduling cannot guarantee delay fairness, which is concluded from the fact that with a heavy traffic load, the worst user suffers from an extremely long delay while the mean delay of the best user is very short. However, M-LWDF, EXP, and MDU, all of which have the maximum stability region, can maintain fairness in terms of delay performance.

Figure 3.3 also depicts the effect of FC. In Section 3.4, we concluded that the FC cannot stabilize the system. It is shown in Figure 3.3 that although the FC enhances the performance of the PF scheduling, the PF-FC does not have the maximum stability region. The effect of FC on the M-LWDF and EXP rules is very small, particularly with a heavy traffic load. However, the FC can boost the performance of the MDU scheduling. When the traffic load is light or moderate, the MDU-FC can reduce the mean delay to half that with the MDU when not using the FC.

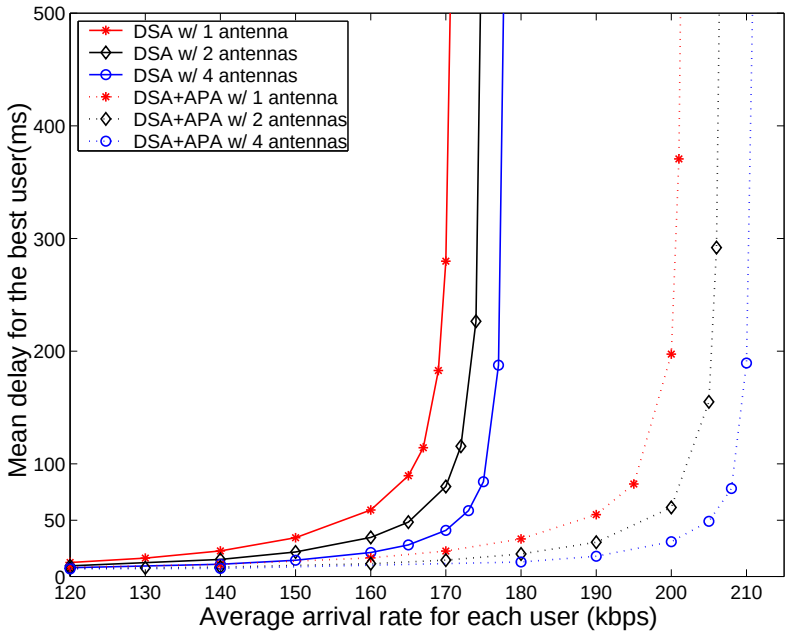
Finally, Figure 3.3 demonstrates that a scheduling scheme with the maximum stability region cannot sufficiently provide good performance. The EXP scheduling has the maximum stability region, but its performance is still poor compared to the M-LWDF and the MDU approaches. This is due to the mechanism of the EXP rule. If one user has a larger delay than others, the weight of this user becomes very large because of the exponential function used in the weight, and then this user may occupy all of the subcarriers with high probability. Because the frequency-selective fading is present, assigning the whole bandwidth to one user is less efficient. Therefore, unlike single-carrier networks, aggressive weight assignments hurt the efficiency in OFDM networks. This is a big difference in designing scheduling between single and multiple carrier networks. It is shown that the multichannel version of M-LWDF works well. However, since the MDU policy uses the average queue lengths (delays) as the weights, which is a more moderate way, the MDU policy can allocate resources more efficiently. Figure 3.3 shows that the MDU-FC outperforms the other schemes.

3.7.2 Improvement of Delay Transmit Diversity and Adaptive Power Allocation

In this subsection, we present the simulation results when delay transmit diversity and power allocation are used. The MDU scheduling with FC is applied in the simulation.



(a) The average waiting time for the worst user



(b) The average waiting time for the best user

Figure 3.4. Delay performance of MDU-FC with delay transmit diversity and adaptive power allocation

We let one delay tap in the delay transmit diversity be 1 μs. Figure 3.4 shows that the joint DSA and APA scheme exhibits a substantial improvement on the throughput-delay performance. Taking advantage of transparency and simplicity, the delay transmit diversity can boost the performance of the scheduling schemes based on DSA as well as those having joint DSA and APA.

3.8 Summary

We have investigated joint channel- and queue-aware multichannel scheduling in OFDM networks from several important aspects. Based on utility functions with respect to average waiting times, we proposed MDU scheduling, which can be implemented by an on-line algorithm without knowledge of the statistical information about the channels and arrival traffic. Since the stability issue of scheduling is essential for QoS provisioning, we characterized the maximum stability region, which can reach the interior of the ergodic physical-layer capacity region. Through concise proofs, we showed that with very few conditions on the scheduling schemes, the scheduling schemes can achieve the maximum stability region. To deal with environments with insufficient scattering or strong light-of-sight components, we proposed using delay transmit diversity to induce the randomness in the frequency domain. In the simulation, we compared several scheduling schemes, and showed that MDU-FC scheduling has better throughput-delay performance than other scheduling schemes, and that the combination of scheduling and power allocation can significantly improve the performance further.

CHAPTER 4

UTILITY-BASED GENERALIZED QOS SCHEDULING FOR HETEROGENEOUS TRAFFIC

We developed the MDU scheduling with the help of channel and queue state information to enhance spectral efficiency and guarantee QoS in Chapter 3, in which, however, we emphasized its theoretical framework, such as queueing system stability. In this chapter, we apply the MDU scheduling to allocate resources for QoS differentiation for different applications. We also present comprehensive simulation results that consider multiple traffic types, including packet-switched voice, streaming, and best-effort traffic. The simulation results demonstrate that the MDU scheduling is a generalized QoS scheduling algorithm that is able to efficiently allocate resources for heterogeneous traffic with diverse QoS requirements. It substantially outperforms the multichannel version of a combination of M-LWDF [8] and PF scheduling [67, 76], called M-LWDF-PF scheduling.

4.1 Introduction

Guaranteeing QoS for multiple types of traffic is challenging to resource allocation and scheduling, especially for wireless data networks [4]. Traditionally, the main idea of QoS provisioning is to reserve resources so as to ensure that certain subjective or objective performance measures are met. In terms of scheduling, *generalized processor sharing* [50] (GPS)-based scheduling schemes, such as *weighted fair queueing* (WFQ) [9], and priority queueing are usually proposed for worst-case throughput and delay guarantee [12, 44, 77]. Obviously, their major drawback is that they cannot improve capacity since no CSI is used.

Currently, channel-aware or opportunistic scheduling has received much attention since it can exploit the variations of wireless fading channels to improve the spectral efficiency [11, 42]. Although proper fairness can be maintained by it, channel-aware scheduling is mainly suitable to best-effort applications but not efficient for delay-sensitive applications.

M-LWDF [6, 8], which makes scheduling decisions based on the current channel conditions and the states of the queues, is proposed for delay-sensitive applications with a QoS requirement that is defined as follows:

$$\mathbb{P}\{W_i > T_i\} \leq \delta_i, \tag{4.1}$$

where W_i is a packet delay for user i, and parameters T_i and δ_i are the delay threshold and the maximum probability of exceeding it, respectively. Its multichannel version is proposed in Section 3.2. To meet QoS differentiation on delay performance, the M-LWDF maps the QoS requirement (4.1) to a scheduling weight

$$a_i = -\frac{\log \delta_i}{T_i}, \tag{4.2}$$

which is based on the results of large deviations. The M-LWDF scheduling is widely used in 1xEV-DO/DV for scheduling delay-sensitive traffic. In addition, it has the maximum stability region. However, the M-LWDF scheduling cannot well handle more complicated QoS requirements and heterogenous traffic.

In the MDU scheduling, the QoS requirements of each user are described by its utility function. From the application view, the MDU scheduling captures the essence of QoS levels with a detail sufficient to predict subjective quality of users. From the network view, it provides the simplicity to enable monitoring and control mechanisms for guaranteeing QoS. By maximizing the total utility within the network, the MDU scheduling establishes a simple, automatic mechanism that can simultaneously improve the spectral efficiency and provide right incentives to ensure that all applications can receive their required QoS.

4.2 MDU Scheduling for Heterogeneous Traffic

In this section, we show how to employ the MDU scheduling for a mixture of delay-sensitive and best-effort traffic by designing utility functions according to the QoS requirements. Note that the MDU scheduling implements the FC in this chapter.

4.2.1 Mechanisms of MDU Scheduling for Diverse QoS Requirements

To apply the MDU scheduling, we need to design utility functions with respect to average waiting time W for the corresponding QoS requirements. Since the marginal utility functions are proportional to the scheduling weights, the marginal utility functions, the $U_i'(\cdot)$'s, play a crucial role in scheduling. Therefore, we directly design the marginal utility functions rather than the utility functions in this section.

We design marginal utility functions based on both certain objective and subjective performance criteria. The objective consideration is the system stability, which is studied in Section 3.4. One of results is that conditions (3.22)-(3.24) can make the MDU scheduling stabilize the queueing system. From a system perspective, a significant different between delay-sensitive and best-effort applications is that the incoming rate of a delay-sensitive stream is usually determined by its source, but the data rate of a best-effort connection is controlled by its transport layer according to the level of network congestion [28]. From a subject perspective, best-effort applications have no specific QoS requirements. Based on these two reasons, *the core idea of designing marginal utility functions is to let the marginal utility functions of delay-sensitive traffic satisfy conditions (3.22)-(3.24), but make the marginal utility functions of best-effort traffic bounded.* Assume that connections 1 to M_1 are delay-sensitive, connections $M_1 + 1$ to M ($M_1 < M$) are best-effort. Their corresponding incoming rate are the λ_i's. It follows from the design that

$$\lim_{W\to\infty} \frac{U_i'(W)}{U_j'(W)} = 0,\ i \in \{M_1 + 1, M_1 + 2, \cdots, M\} \text{ and } j \in \{1, 2, \cdots, M_1\}. \tag{4.3}$$

The above equation means that the MDU scheduling can sense the level of network congestion. If the network is congested, best-effort connections hardly obtain resources to transmit packets according to (4.3). If rate vector $[\lambda_1, \lambda_2, \cdots, \lambda_{M_1}, 0, 0, \cdots, 0]^T$ is located within the ergodic capacity region $\tilde{\mathcal{C}}$, the above design makes all of delay-sensitive connections stable, which comes directly from the results of Corollary 3.1. Therefore, the MDU scheduling does not allow those best-effort connections to affect the stability of delay-sensitive connections. If the network load is low, the scheduler can automatically assign more resources to those best-effort connections. The more specific design of the marginal utility functions is

based on the subjective performance criteria of certain applications. Section 4.2.2 shows the details.

4.2.2 Marginal Utility Functions for MDU Scheduling

In this section, we design the marginal utility functions based on the corresponding required QoS for packet-switched voice, streaming, and best-effort traffic.

4.2.2.1 Delay-sensitive Applications

For a delay-sensitive application, we set a threshold for the marginal utility function that depends on the characteristics of the application. When the average waiting time is less than the threshold, the marginal utility increases with a small order. When the average waiting time is beyond the threshold, the marginal utility increases with a relatively high order.

For packet-switched voice or *voice over IP* (VoIP), the end-to-end delay is usually required less than 100 ms [1]. Since there are other delay factors besides the delay resulting from wireless scheduling, we set the marginal utility function for voice as follows:

$$|U'_V(W)| = \begin{cases} W & W \leq 25\text{ms} \\ W^{1.5} - 25^{1.5} + 25 & W > 25\text{ms}, \end{cases} \tag{4.4}$$

where the threshold, 25 ms, comes from one-forth of 100 ms.

Good-quality streaming transmission needs end-to-end delay between 150-400 ms. We choose the following marginal utility function for streaming traffic.

$$|U'_S(W)| = \begin{cases} W^{0.6} & W \leq 100\text{ms} \\ W - 100 + 100^{0.6} & W > 100\text{ms}, \end{cases} \tag{4.5}$$

where the threshold, 100 ms, comes from one-forth of 400 ms. Obviously, marginal utility functions (4.4) and (4.5) both satisfy conditions (3.22)-(3.24) .

4.2.2.2 Best-Effort Applications

Since best-effort traffic is not delay-sensitive, the utility function with respect to average waiting time is not enough sufficient to describe the performance of this traffic. From a

point of view of scheduling weights, however, we can still give the marginal utility function in terms of average waiting time. For example, the marginal utility function is given by

$$|U'_D(W)| = \begin{cases} W^{0.5} & W \leq 100\text{ms} \\ 100^{0.5} & W > 100\text{ms}. \end{cases} \tag{4.6}$$

Please note a long average waiting time means network congestion. Compared to (4.4) and (4.5), the marginal utility function for best-effort traffic (4.6) lets the scheduling weights be bounded, whereas delay-sensitive applications set higher scheduling weights according to (4.4) and (4.5). In this simulation, we intend to know the maximum throughput that best-effort traffic can obtain. Thus, we fix $|U'_D(W)|$ to the maximum value, $100^{0.5}$. Actually, the MDU scheduling for the best-effort traffic becomes the PF scheduling, which is known to be well applicable to best-effort traffic in Chapter 2.

4.3 Simulation

In this section, we design appropriate simulations that take into account the impacts of different traffic types and average SNR values on scheduling performance.

4.3.1 Simulation Conditions

For comparison, we assume that the number of each traffic type is an even integer. For each type of traffic, half of users have the same average SNR of 15 dB, and we call them good users; the rest have the same average SNR of 8 dB, we call them bad users. In the simulation, each bad user's channel suffers multipath Rayleigh fading with the delay profile of Channel B for outdoor to indoor and pedestrian environments in [54], and each user is assumed to be stationary or slowly moving so that the maximum Doppler shift is 10 Hz. Each good user experiences Rician fading whose delay profile and Doppler shift are the same as those of bad users' channels. The Rician factor is 0.5. In the OFDM network, there are 256 subcarriers in a total channel bandwidth of 2.048 MHz. These 256 subcarriers are grouped into 32 clusters, each of which can be dynamically assigned to a user during a time slot. Let the acceptable BER be 10^{-5} for rate adaptation since data transmission is sensitive to error. Assume that a set of achievable transmission rates in bits/sec per Hz is

{0,1/2,1,2,3,4}. In practice, we can use 1/2-rate channel coding and a series of modulation schemes including BPSK, QPSK, 16-QAM, 64-QAM, as well as 256-QAM to achieve the above feasible rates.

We consider three types of traffic: packet-switched voice, streaming, and best-effort traffic. The traffic model for voice traffic is the on-off voice activity model with exponentially distributed duration of voice spurts and gaps [60]. The average talk spurt is 1.00 s, and the average silent interval is 1.35 s. Within each talk spurt interval, a 32 kbps digital voice coding is assumed. The streaming traffic is simulated according to the model in [19]. The duration of each state is exponentially distributed with mean 160 ms. The data rate in each state is generated according to a truncated exponential distribution in which the minimum, maximum, and average data rates are 64, 256, and 180 kbps, respectively. As mentioned before, we only care about the maximum throughput of best-effort traffic in this simulation and fix its scheduling weights. Therefore, we apply a full-buffer model to best-effort traffic. In the full-buffer model, there are infinite data packets in the queues. Although this model may not be realistic, it can obtain the maximum achievable throughput for best-effort traffic.

For M-LWDF-PF scheduling, the weights for delay-sensitive applications can be calculated by (4.2). The weights in simulation are list in the following table.

Table 4.1. Scheduling weights for M-LWDF-PF

	Voice	Streaming	Best-effort
T_i (ms)	100	400	–
δ_i	5%	5%	–
Weight	13	3.25	0.26

4.3.2 Simulation Results

We design three experiments in the simulation and compare the performance of the MDU scheduling and that of the M-LWDF-PF scheduling. The performance of delay-sensitive traffic is evaluated in terms of 95th percentile delay, and that of best-effort traffic is measured in terms of average throughput. We focus on discussing the properties of the MDU scheduling at first.

4.3.2.1 Increase of voice users

In this experiment, we fix the numbers of streaming and best-effort users to be 14 and 20, respectively, and increase the number of voice users. It is seen from Figure 4.1 that as the number of voice users increases, the throughput of best-effort traffic decreases apparently; the delay for streaming users increases slightly. However, there is only a very little rise in the delay for voice and streaming users in the system employing the MDU scheduling.

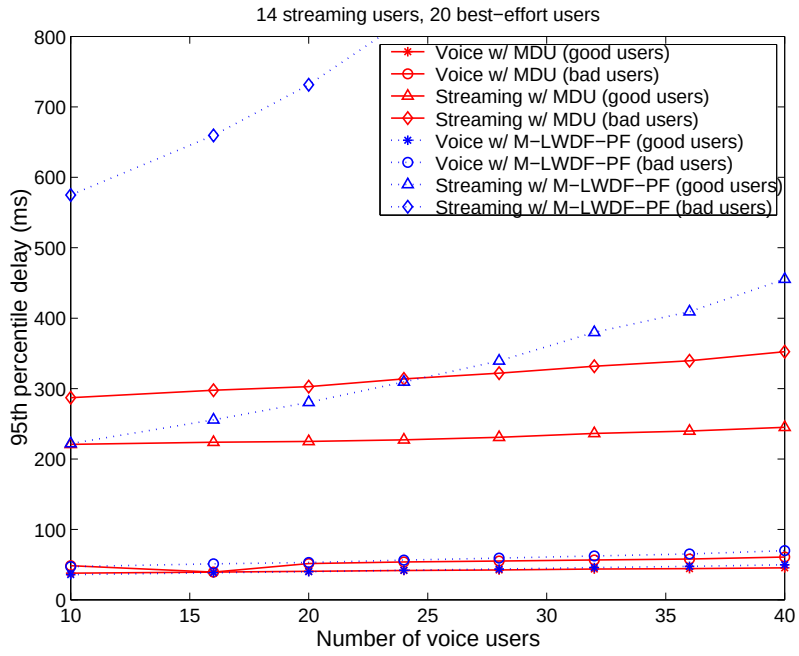
4.3.2.2 Increase of streaming users

In this experiment, we fix the numbers of voice and best-effort users both to be 20 and increase the number of streaming users. Since the average data rate of a streaming link is as large as 180 kbps, we can clearly see the performance in both less-congested and congested situations in Figure 4.2. When the network is less-congested (the number of streaming users does not exceed 16), the MDU scheduling can maintain high-quality delay performance for those delay-sensitive applications and provide a high data rate for the best-effort users. When the network is congested, e.g. in the 20-streaming-user case, the throughput for best-effort users becomes extremely small, and the delay for streaming users has a dramatical increase. However, the performance of voice users is still very good.

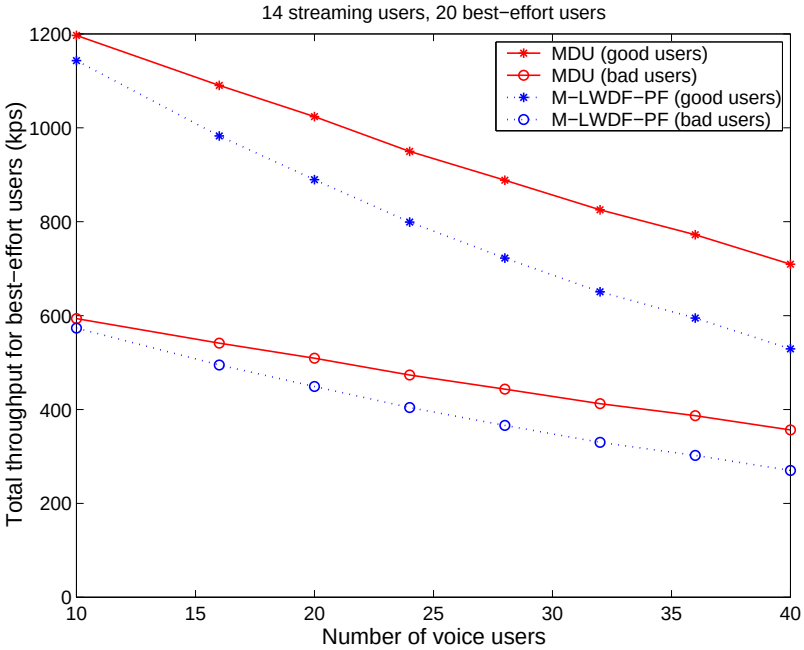
4.3.2.3 Increase of best-effort users

In the last experiment, we fix the numbers of voice and streaming users to be 20 and 10, respectively, and increase the number of best-effort users. It is seen from Figure 4.3 that as the number of best-effort users increases, the performance of voice and streaming users is maintained very well with the MDU scheduling, and the throughput for best-effort increases, which results from multiuser diversity.

Therefore, we can in these three experiments see the excellent mechanisms of the MDU scheduling: high spectral efficiency by taking advantage of knowledge of CSI and good diverse QoS provisioning by exploiting utility functions. We also compare the MDU with the M-LWDF-PF in Figures 4.1-4.3. Note that the M-LWDF scheduling is also a scheduling

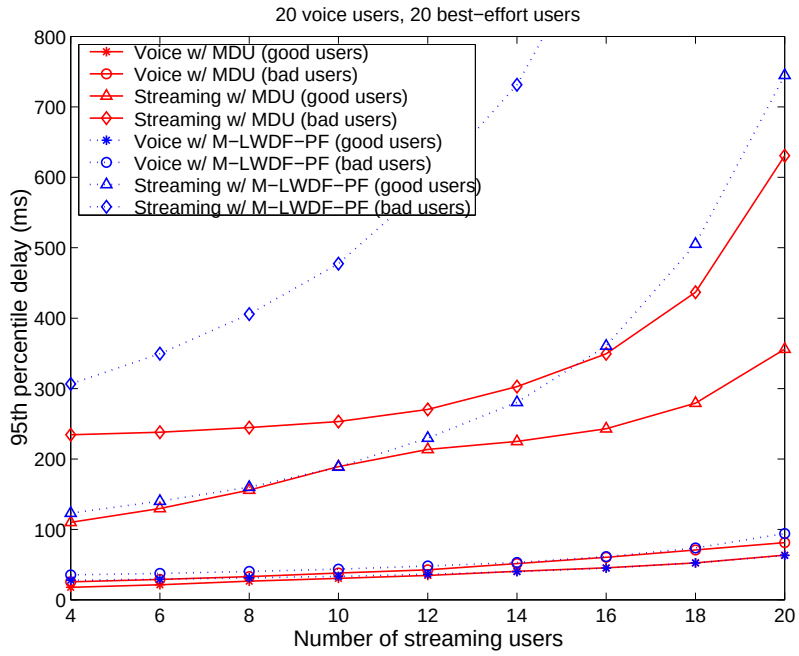


(a) 95th percentile delay for voice and streaming traffic

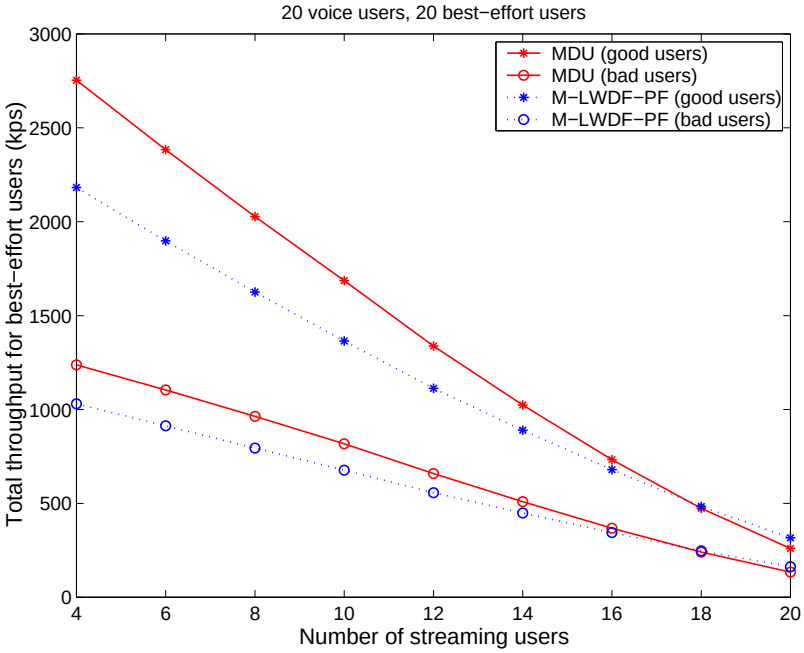


(b) Average total throughput for best-effort traffic

Figure 4.1. Heterogeneous traffic performance versus the number of voice users

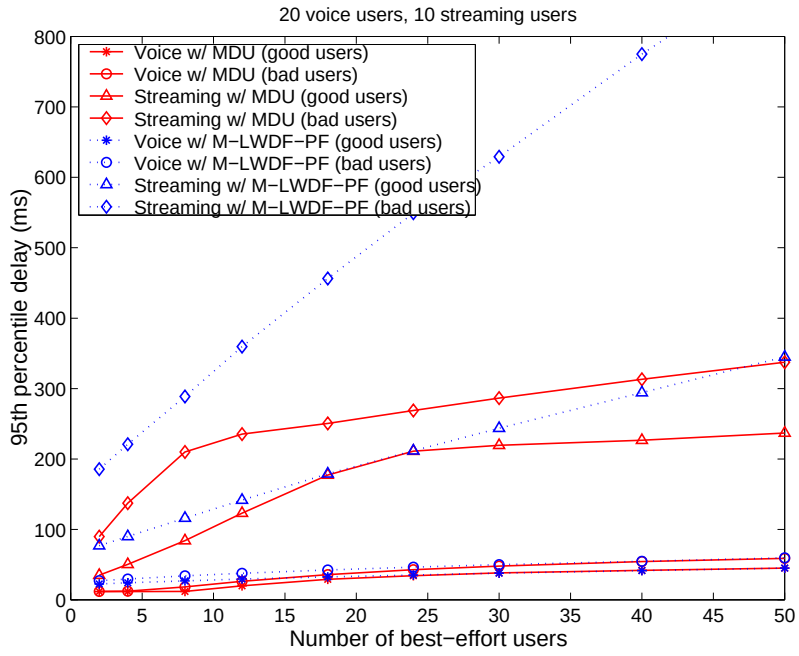


(a) 95th percentile delay for voice and streaming traffic

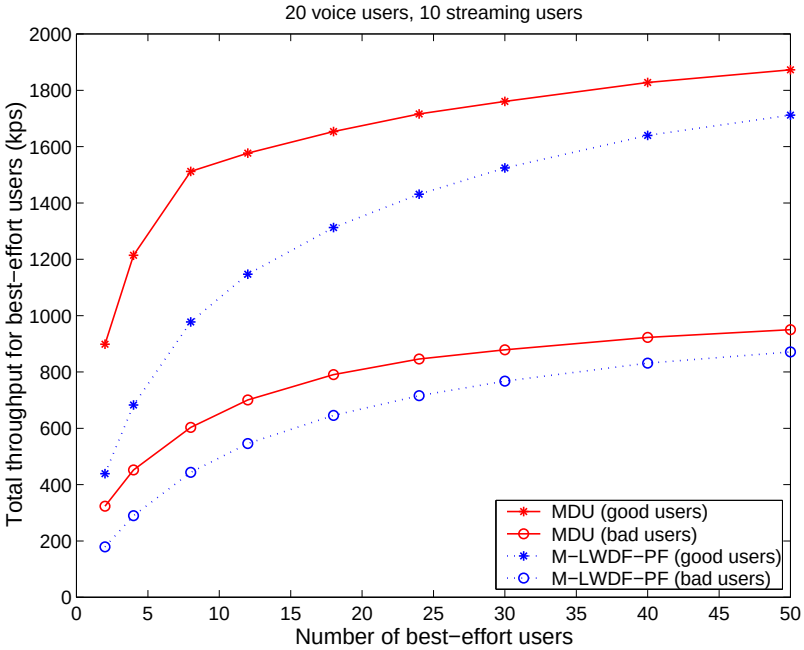


(b) Average total throughput for best-effort traffic

Figure 4.2. Heterogeneous traffic performance versus the number of streaming users



(a) 95th percentile delay for voice and streaming traffic



(b) Average total throughput for best-effort traffic

Figure 4.3. Heterogeneous traffic performance versus the number of best-effort users

scheme that can adjust resource allocation according to users' channel and queue state information and have the maximum stability region. All experiments show that both scheduling schemes offer similar delay performance for the voice users, and that in most of the cases, the MDU scheduling provides considerably smaller delays for streaming traffic than the M-LWDF-PF while the MDU allows best-effort users to achieve higher throughput than the M-LWDF-PF at the same time. This is mainly because the MDU scheduling can more appropriately capture required QoS compared to other scheduling schemes.

4.4 Summary

By simulation, we have demonstrated that the MDU scheduling can effectively handle multiple traffic types with diverse QoS requirements and substantially outperforms the multichannel version of the M-LWDF-PF scheduling. The MDU scheduling benefits from the awareness of channel quality and queue information, traffic multiplexing, and resource regulation through utility functions, which appropriately capture the QoS requirements of specific traffic. In addition, the MDU scheduling has a very simple QoS architecture. It does not need statistical information about incoming traffic, and its implementation complexity is also low. Therefore, the MDU scheduling is an attractive solution for IP-based wireless data networks.

CHAPTER 5

ASYMPTOTIC PERFORMANCE ANALYSIS FOR CHANNEL-AWARE SCHEDULING

To obtain the multiuser diversity gain, adaptive modulation and channel-aware scheduling must be used. However, the channel variance and the opportunistic nature of channel-aware scheduling make throughput analysis very difficult. An asymptotic analysis of SNR for multiuser diversity is presented in [76]. Capacity analyses for Rayleigh and Nakagami fading channels are addressed in [81] and [14], respectively. However, the results of those capacity analyses are too complicated to get insights.

In this chapter, we provide an asymptotic performance analysis of channel-aware packet scheduling based on extreme value theory, including throughput an delay analysis for both sing-carrier and multicarrier networks. In Section 5.1, we briefly describe the main results of extreme value theory used in this chapter. In Section 5.2, we propose an asymptotic analysis of throughput of single-carrier systems with channel-aware scheduling. we first address the average throughput of systems with a homogeneous average SNR and obtain its asymptotic expression. Compared to the exact throughput expression, the asymptotic one, which is applicable to a broader range of channel fading distributions, is more concise and easier to get insights. Furthermore, we confirm the accuracy of the asymptotic results by numerical simulation. For a system with heterogeneous SNRs, normalized-SNR-based scheduling need to be used for fairness. We also investigate the asymptotic average throughput of the normalized-SNR-based scheduling and prove that the average throughput in this case is less than that in the homogeneous case with a power constraint. In Section 5.3, we provide a closed-form asymptotic average packet delay analysis for single-carrier networks exploiting multiuser diversity. In Section 5.4, asymptotic analysis of throughput and delay is extended into multicarrier networks. The asymptotic analysis for mean packet delay demonstrates that the multiuser diversity gain in multicarrier networks is not limited by slow fading as

in single-carrier networks.

5.1 Extreme Value Theory

Extreme value theory deals with asymptotic distributions of extreme values, such as maxima or minima. It can be used to analyze the performance of channel-aware scheduling approaches. In this section, we will briefly introduce the major results of extreme value theory [18, 22] that are used in the analysis.

Let $\xi_1, \xi_2, \cdots, \xi_M$ be *independently identically distributed* (i.i.d.) random variables with common distribution function $F(x)$. We are interested in the distribution of the maximum, $Z_M = \max_{i \in \mathcal{M}} \xi_i$ as $M \to \infty$. The *cumulative distribution function* (cdf) of the maximum, $H_M(x)$, is given by

$$H_M(x) = \mathbb{P}\{Z_M \leq x\} = F^M(x)$$

When $M \to \infty$, we have

$$F^M(x) \to \begin{cases} 1 & \text{if } F(x) = 1, \\ 0 & \text{if } F(x) = 0, \end{cases}$$

which means that the limiting distribution is degenerate at either 0 or 1. In order to avoid this degeneration, we look for such normalizing constants a_M and b_M depending on M that

$$\begin{aligned} \lim_{M \to \infty} H_M(a_M + b_M x) &= \lim_{M \to \infty} \mathbb{P}\{\frac{Z_M - a_M}{b_M} \leq x\} \\ &= \lim_{M \to \infty} F^M(a_M + b_M x) \\ &= H(x) \end{aligned}$$

where $H(x)$ is a limiting *non-degenerate* distribution function. We also say that $\frac{Z_M - a_M}{b_M}$ converges in the sense of distribution in this case. An important result about limiting distribution is described as follows.

Let $\xi_1, \xi_2, \cdots, \xi_M$ be i.i.d. random variables with distribution function $F(x)$, and $Z_M = \max_{i \in \mathcal{M}} \xi_i$. If there exist constants $a_M \in \mathbb{R}$, and $b_M > 0$, and some non-degenerate distribution function H such that the distribution of $\frac{Z_M - a_M}{b_M}$ converges to H, then H belongs to one of the three standard extreme value distributions: Frechet, Webull, and Gumbel distributions.

It is very interesting that there are only three possible non-degenerate limiting distributions for maxima. The distribution function $F(x)$ determines the exact limiting distribution. Thus, if a distribution function $F(x)$ results in one limiting distribution for extremes, we say that $F(x)$ belongs to the domain of attraction of this limiting distribution. Next, we will introduce a sufficient condition for a distribution function $F(x)$ to belong to the domain of attraction of the Gumbel distribution.

Lemma 5.1 *Let $\xi_1, \xi_2, \cdots, \xi_M$ be i.i.d. random variables with distribution function $F(x)$. Define $\omega(F) = \sup\{x : F(x) < 1\}$. Assume that there is a real number x_1 such that, for all $x_1 \leq x < \omega(F)$, $f(x) = F'(x)$ and $F''(x)$ exist and $f(x) \neq 0$. If*

$$\lim_{x \to \omega(F)} \frac{d}{dx} \left[\frac{1 - F(x)}{f(x)} \right] = 0, \tag{5.1}$$

then there exist sequences a_M and $b_M > 0$ such that, as $M \to \infty$, $\dfrac{Z_M - a_M}{b_M}$ uniformly converges in distribution to a normalized Gumbel (maxima) random variable. The normalizing constants a_M and b_M can be chosen as

$$a_M = F^{-1}\left(1 - \frac{1}{M}\right),$$
$$b_M = F^{-1}\left(1 - \frac{1}{Me}\right) - F^{-1}\left(1 - \frac{1}{M}\right),$$

where $F^{-1}(x) = \inf\{y : F(y) \geq x\}$.

For a random variable Z with the normalized Gumbel distribution for maxima, $\exp[-\exp(-x)]$, $-\infty < x < \infty$, it follows that

$$\mathbb{E}\{Z\} = E_0,$$
$$\mathbb{V}ar\{Z\} = \frac{\pi^2}{6},$$

where $E_0 = 0.5772\cdots$ is the Euler constant [22].

In this chapter, we intend to calculate the average throughput; thus, mean convergence is used extensively. However, convergence in distribution cannot generally guarantee mean convergence. [51] established the relation between convergence in distribution and moment convergence, which is stated in the following lemma.

Lemma 5.2 *If* $\dfrac{Z_M - a_M}{b_M}$ *converges in distribution to a random variable* Z *that has a non-degenerate distribution function, and if* $\mathbb{E}\{[(Z_M)^-]^p\} < \infty$ *for any positive real number* p*, where* $(x)^- = -x,\ x < 0,\ \ = 0$*, otherwise, then*

$$\lim_{M\to\infty} \mathbb{E}\left(\frac{Z_M - a_M}{b_M}\right)^p = \mathbb{E}\{Z^p\},$$

provided $\mathbb{E}|Z|^p < \infty$*.*

Obviously, convergence in distribution for the maximum of *nonnegative* random variables results in moment convergence.

Lemmas 5.1 and 5.2 can be restated for minima by consider $(-\xi_i)$ instead of ξ_i. We give below the result about the asymptotic distribution of the minimum of i.i.d. random variables. Let $W_M = \min_{i\in\mathcal{M}} \xi_i$.

Lemma 5.3 *Let* $\xi_1, \xi_2, \cdots, \xi_M$ *be i.i.d. random variables with distribution function* $F(x)$*. Define* $\alpha(F) = \inf\{x : F(x) > 0\}$*. Assume that there is a real number* x_1 *such that, for all* $\alpha(F) \le x < x_1$*,* $f(x) = F'(x)$ *and* $F''(x)$ *exist and* $f(x) \neq 0$*. If*

$$\lim_{x\to\alpha(F)} \frac{d}{dx}\left[\frac{F(x)}{f(x)}\right] = 0, \tag{5.2}$$

then there exist sequences c_M *and* $d_M > 0$ *such that, as* $M \to \infty$*,* $\dfrac{W_M - c_M}{d_M}$ *uniformly converges in distribution to a normalized Gumbel (minima) random variable. The normalizing constants* c_M *and* d_M *can be chosen as*

$$c_M = F^{-1}\left(\frac{1}{M}\right),$$
$$d_M = F^{-1}\left(\frac{1}{M}\right) - F^{-1}\left(\frac{1}{Me}\right).$$

For a random variable W with the normalized Gumbel distribution for minima, $1-\exp[-\exp(-x)]$, $-\infty < x < \infty$, it follows that

$$\mathbb{E}\{W\} = -E_0,$$
$$\mathbb{V}ar\{W\} = \frac{\pi^2}{6}.$$

With extreme value theory, we can study the asymptotic performance of channel-aware scheduling.

5.2 Asymptotic Throughput Analysis of Single-Carrier Networks

In this section, we focus on asymptotic throughput analysis for single-carrier networks with channel-aware scheduling.

5.2.1 System Model

Consider a shared downlink channel of a single-carrier system with a bandwidth B and M users. The downlink channel is time-slotted, and each time slot can adaptively be assigned to a user. It is assumed that the base station knows the CSI of each user, and that continuous rate adaptation is applied in the downlink channel. Therefore, the transmission data rate, R, depends on the current SNR, Γ. The relationship can be written as [23]

$$R = B \log_2(1 + \beta\Gamma), \tag{5.3}$$

where β is a constant related to the targeted BER and the used modulation and coding techniques.

First, we assume that the all users experience statistically independent identical fading processes. The *max-sum-capacity* (MSC) scheduling rule [66, 76] is used in the system. The MSC rule is a channel-aware scheduling scheme that maximizes the total throughput in the system and works well in the homogeneous system. It assigns the channel to the user with the best channel condition on each time slot, which is described as

$$m = \arg\max_{i \in \mathcal{M}} \{\Gamma_i\}, \tag{5.4}$$

where $\mathcal{M} = \{1, 2, \cdots, M\}$, and Γ_i is the SNR of user i.

We also consider the heterogeneous case, in which different users have different average SNR values due to various path losses. For the purpose of fairness, the normalized-SNR-based scheduling [81] is used. This scheduling rule makes decisions based on the normalized SNR rather than the absolute SNR values, which is expressed as

$$m = \arg\max_{i \in \mathcal{M}} \{\frac{\Gamma_i}{\gamma_i}\}, \tag{5.5}$$

where γ_i is the average SNR of user i; that is, $\mathbb{E}\{\Gamma_i\} = \gamma_i$. It is obvious that the normalized-SNR-based scheduling is equivalent to the MSC scheduling in the homogeneous system.

5.2.2 Throughput Analysis for Rayleigh Fading

In this section, we will analyze the throughput performance of multiuser diversity for Rayleigh fading channels with the same average SNR γ_0. The cdf of the SNR for Rayleigh fading can be expressed as

$$F_\Gamma(\gamma) = 1 - \exp(-\frac{\gamma}{\gamma_0}). \tag{5.6}$$

5.2.2.1 Exact Analysis

According to the MSC scheduling, the base station schedules the user with the strongest channel condition. Therefore, the effective SNR at the transmitter, Γ_{eff}, is given by

$$\Gamma_{\text{eff}} = \max_{i \in \mathcal{M}} \Gamma_i, \tag{5.7}$$

and its distribution is

$$\begin{aligned} F_{\Gamma_{\text{eff}}}(\gamma) &= \mathbb{P}\{\Gamma_1 < \gamma, \Gamma_2 < \gamma, \cdots, \Gamma_M < \gamma\} \\ &= \left(1 - e^{-\gamma/\gamma_0}\right)^M \end{aligned}$$

By taking derivative, the pdf of Γ_{eff} can be obtained as

$$\begin{aligned} f_{\Gamma_{\text{eff}}}(\gamma) &= \frac{d}{d\gamma} F_{\Gamma_{\text{eff}}}(\gamma) \\ &= M(1 - e^{-\gamma/\gamma_0})^{M-1} \frac{e^{-\gamma/\gamma_0}}{\gamma_0}. \end{aligned} \tag{5.8}$$

Using (5.8), we calculate the average SNR when the MSC scheduling is used as

$$\begin{aligned} \mathbb{E}\{\Gamma_{\text{eff}}\} &= \int_0^\infty \gamma \cdot f_{\Gamma_{\text{eff}}}(\gamma)\, d\gamma \\ &= \gamma_0 \sum_{i=1}^{M} \frac{1}{i}. \end{aligned} \tag{5.9}$$

The throughput of the MSC scheduling is expressed as

$$R_{\text{total}} = B \log_2(1 + \beta \Gamma_{\text{eff}});$$

hence the average throughput is

$$\mathbb{E}\{R_{\text{total}}\} = B \int_0^\infty \log_2(1 + \beta\gamma) f_{\Gamma_{\text{eff}}}(\gamma)\, d\gamma. \tag{5.10}$$

To obtain a closed-form result of $\mathbb{E}\{R_{\text{total}}\}$, rewriting (5.8) by using the binomial expansion as

$$f_{\Gamma_{\text{eff}}}(\gamma) = \frac{M}{\gamma_0} \sum_{i=0}^{M-1} (-1)^i \begin{pmatrix} M-1 \\ i \end{pmatrix} e^{-\frac{(1+i)\gamma}{\gamma_0}}, \tag{5.11}$$

where

$$\begin{pmatrix} M-1 \\ i \end{pmatrix} = \frac{(M-1)!}{(M-i-1)!\ i!}.$$

Substituting (5.11) into (5.10), we obtain

$$\mathbb{E}\{R_{\text{total}}\} = \frac{M}{\ln 2} \sum_{i=0}^{M-1} (-1)^{i+1} \begin{pmatrix} M-1 \\ i \end{pmatrix} \frac{e^{\frac{1+i}{\gamma_0}}}{i+1} E_i(-\frac{1+i}{\gamma_0}) \tag{5.12}$$

with

$$E_i(-x) = E_0 + \ln(x) + \sum_{i=1}^{\infty} \frac{(-1)^i x^i}{i!\ \ i}.$$

As seen above, the exact analysis of throughput analysis is very complicated, and the exact result lacks insights. Therefore, in the rest of the chapter, we will provide the simple results through asymptotic analysis.

5.2.2.2 Asymptotic Analysis

First, we study the asymptotic distribution for the effective SNR Γ_{eff} in (5.7). The exponential distribution leads to

$$\frac{1 - F_\Gamma(\gamma)}{f_\Gamma(\gamma)} = \gamma_0.$$

As a result, it follows that

$$\frac{d}{d\gamma}\left[\frac{1 - F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right] = 0, \text{ for } \gamma > 0.$$

According to the results of extreme value theory in Section 5.1, the exponential distribution is in the domain of attraction of the Gumbel distribution, and

$$a_M = \gamma_0 \ln M,$$

$$b_M = \gamma_0.$$

According to Lemma 5.2, as $M \to \infty$,

$$\frac{\mathbb{E}\{\Gamma_{\text{eff}}\} - \gamma_0 \ln M}{\gamma_0} \to E_0.$$

With a large M, therefore,

$$\mathbb{E}\{\Gamma_{\text{eff}}\} \approx \gamma_0(\ln M + E_0). \tag{5.13}$$

It is shown in [82] that

$$\frac{1}{2(M+1)} < \sum_{i=1}^{M} \frac{1}{i} - (\ln M + E_0) < \frac{1}{2M},$$

which implies that the difference between the exact value (5.9) and the asymptotic value (5.13) is very small even for a small M.

We can also use the results of extreme value theory in Section 5.1 to obtain an asymptotic analysis for throughput $R_{\text{total}} = \max_{i \in \mathcal{M}} R_i$. Let

$$R = T(\Gamma) \triangleq B \log_2(1 + \beta\Gamma).$$

Since $T(\Gamma)$ is a monotonic increasing function of Γ, the distribution of data rate R is given by

$$F_R(r) = F_\Gamma(T^{-1}(r)),$$

where $T^{-1}(r) = \dfrac{2^{\frac{r}{B}} - 1}{\beta}$. For the Rayleigh fading channel, we have

$$\begin{aligned} \frac{1 - F_R(r)}{f_R(r)} &= \frac{1 - F_\Gamma(T^{-1}(r))}{f_\Gamma(T^{-1}(r))\,(T^{-1})'(r)} \\ &= \frac{\beta\gamma_0}{2^{\frac{r}{B}} - 1}. \end{aligned} \tag{5.14}$$

Equation (5.14) results in

$$\lim_{r \to \infty} \frac{d}{dr}\left[\frac{1 - F_R(r)}{f_R(r)}\right] = 0. \tag{5.15}$$

According to the results of extreme value theory in Section 5.1, therefore, the maximum throughput, $R_{\text{total}} = \max_{i \in \mathcal{M}} R_i$, asymptotically behaves as a Gumbel random variable,

$a_M + b_M Z$, where Z is a normalized Gumbel random variable, and

$$a_M = B\log_2(1+\beta\gamma_0 \ln M),$$
$$b_M = B\log_2\left(\frac{1+\beta\gamma_0(1+\ln M)}{1+\beta\gamma_0 \ln M}\right).$$

Moreover, as $M \to \infty$,

$$\frac{E\{R_{\text{total}}\} - a_M}{b_M} \to \mathbb{E}\{Z\} = E_0.$$

Thus, when M is large, the average throughput is given by

$$\begin{aligned}\mathbb{E}\{R_{\text{total}}\} &\approx a_M + E_0 b_M \\ &= B\log_2(1+\beta\gamma_0 \ln M) + E_0 \cdot B\log_2\left(\frac{1+\beta\gamma_0(1+\ln M)}{1+\beta\gamma_0 \ln M}\right), \end{aligned} \tag{5.16}$$

where $\ln M$ is called multiuser diversity gain [76]. In contrast to (5.12), (5.16) provides a very simple approximation for the average throughput. The numerical results in Section 5.2.5 will shows that this approximation is very accurate.

Note that as $M \to \infty$, $a_M \to \infty$, and $b_M \to 0$. Therefore, with a large M,

$$\mathbb{E}\{R_{\text{total}}\} \approx B\log_2(1+\beta\gamma_0 \ln M)$$

is a rougher but simpler estimation for the average throughput. For the Rayleigh fading, it is easy to prove (5.15). However, proving (5.15) may be difficult for other fading distributions. We will provide a simple way to do it in the next section.

5.2.3 Throughput Analysis for General Channel Distributions

In previous sections, we have seen that finding the limiting distribution of the maximum throughput is crucial to obtain the asymptotic throughput. In this section, we consider more general cases beyond Rayleigh fading. Mathematically, we study the limiting distribution of the throughput

$$R = T(\Gamma) = B\log_2(1+\beta\Gamma),$$

given a SNR distribution, $F_\Gamma(\gamma)$. The major result is stated in the following theorem for *limiting throughput distribution* (LTD).

Theorem 5.1 *(*LTD Theorem*): Assume that all users' SNRs,* $\{\Gamma_1, \Gamma_2, \cdots, \Gamma_M\}$*, are i.i.d. random variables with a distribution* $F_\Gamma(\gamma)$ *such that* $\omega(F_\Gamma) = \infty$*, and* $f_\Gamma(\gamma) = F'_\Gamma(\gamma)$ *as well as* $F''_\Gamma(\gamma)$ *exist and* $f_\Gamma(\gamma) \neq 0$ *for all* $x_1 \leq x < \infty$*, where* x_1 *is some real number. If*

$$\lim_{\gamma \to \infty} \frac{d}{d\gamma} \left[\frac{1 - F_\Gamma(\gamma)}{f_\Gamma(\gamma)} \right] = 0, \tag{5.17}$$

then the distribution of throughput, $F_R(r) = F_\Gamma(T^{-1}(r))$*, belongs to the domain of the attraction of the Gumbel distribution (maxima). In addition,*

$$a_M = B \log_2 \left(1 + \beta F_\Gamma^{-1}(1 - \frac{1}{M}) \right), \tag{5.18}$$

$$b_M = B \log_2 \left(\frac{1 + \beta F_\Gamma^{-1}(1 - \frac{1}{Me})}{1 + \beta F_\Gamma^{-1}(1 - \frac{1}{M})} \right). \tag{5.19}$$

The proof is shown in Appendix G. the LTD theorem tells us that we do not have to check $F_R(r)$ directly, which is usually very complicated to find its limiting distribution. In addition, Lemma 5.2 leads to

$$\frac{\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} - a_M}{b_M} \to E_0,$$

as $M \to \infty$, where $R_{\text{total}}^{\text{hom}}$ is the total throughput for the homogeneous scenario. For a large M, the average total throughput can be evaluated by using the following expression.

$$\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} \approx a_M + E_0 b_M. \tag{5.20}$$

5.2.3.1 Example

The Nakagami distribution is frequently used to characterize the fading statistics of wireless channels in certain environments. Then, the cdf of the received SNR is given by

$$F_\Gamma(\gamma) = \Gamma_{(m, \frac{m}{\gamma_0})}(\gamma) = \int_0^\gamma \left(\frac{m}{\gamma_0} \right)^m \frac{t^{m-1}}{\Gamma(m)} e^{-\frac{m}{\gamma_0} t} dt, \tag{5.21}$$

where m is called the fading figure, which is defined as the ratio of the total power to the power of fading components, and $\Gamma(m)$ is the gamma function. In this subsection, we use the results of the LTD theorem to study the impact of Nakagami fading on throughput in the system with the MSC scheduling. Applying the results of extreme value theory in

Section 5.1 and letting $u = \frac{m}{\gamma_0}$, we have

$$\begin{aligned}
&\lim_{\gamma\to\infty} \frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right] \\
&= \lim_{\gamma\to\infty} -\frac{[1-F_\Gamma(\gamma)]}{f_\Gamma^2(\gamma)/f_\Gamma'(\gamma)} - 1 \\
&= \lim_{\gamma\to\infty} \frac{1-\int_0^\gamma t^{m-1}e^{-ut}dt}{\frac{\gamma^m e^{-u\gamma}}{u\gamma-m+1}} - 1 \\
&= 0 \quad \text{(by L'Hospital's rule).}
\end{aligned}$$

According to the results of extreme value theory in Section 5.1 and the LTD theorem, both $F_\Gamma(\gamma)$ and $F_R(r)$ belong to the domain of the attraction of the Gumbel distribution. Therefore, the average total throughput for the Nakagami fading can be given by

$$\begin{aligned}
\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} &\approx B\log_2\left(1+\beta F_\Gamma^{-1}(1-\frac{1}{M})\right) + E_0 B\log_2\left(\frac{1+\beta F_\Gamma^{-1}(1-\frac{1}{Me})}{1+\beta F_\Gamma^{-1}(1-\frac{1}{M})}\right) \\
&= B\log_2\left(1+\beta\Gamma^{-1}_{(m,\frac{m}{\gamma_0})}(1-\frac{1}{M})\right) + E_0 B\log_2\left(\frac{1+\beta\Gamma^{-1}_{(m,\frac{m}{\gamma_0})}(1-\frac{1}{Me})}{1+\beta\Gamma^{-1}_{(m,\frac{m}{\gamma_0})}(1-\frac{1}{M})}\right),
\end{aligned} \tag{5.22}$$

where $\Gamma^{-1}_{(m,\frac{m}{\gamma_0})}(\gamma)$ is the inverse incomplete gamma function. Despite no closed form for it, the inverse incomplete gamma function is usually provided in common softwares, such as Matlab and Mathematica.

Actually, besides the Rayleigh and Nakagami distributions, the normal, Rician, and log-normal distributions, which are often used to describe the statistics of wireless channels, belong to the domain of the attraction of the Gumbel distribution [22].

5.2.3.2 Further Properties of Asymptotic Throughput

Note that as $M \to \infty$, $a_M \to \infty$ since $F_\Gamma^{-1}(\gamma) \to \infty$ as $\gamma \to \infty$. In addition, in Appendix H, we prove that

$$\lim_{M\to\infty} \frac{b_M}{a_M} = 0. \tag{5.23}$$

Applying (5.23) to $F_\Gamma(\gamma)$ (a_M and b_M here are related to Γ_{eff} in (5.7)), we have

$$\lim_{M\to\infty} \frac{F_\Gamma^{-1}(1-\frac{1}{Me}) - F_\Gamma^{-1}(1-\frac{1}{M})}{F_\Gamma^{-1}(1-\frac{1}{M})} = 0. \tag{5.24}$$

From (5.24), we have the limit of b_M that is corresponding to the throughput as follows:

$$\begin{aligned}\lim_{M\to\infty} b_M &= \lim_{M\to\infty} B\log_2\left(\frac{1+\beta F_\Gamma^{-1}(1-\frac{1}{Me})}{1+\beta F_\Gamma^{-1}(1-\frac{1}{M})}\right)\\ &= \lim_{M\to\infty} B\log_2\left(\frac{F_\Gamma^{-1}(1-\frac{1}{Me})}{F_\Gamma^{-1}(1-\frac{1}{M})}\right)\\ &= \lim_{M\to\infty} B\log_2\left(\frac{F_\Gamma^{-1}(1-\frac{1}{Me})-F_\Gamma^{-1}(1-\frac{1}{M})}{F_\Gamma^{-1}(1-\frac{1}{M})}+1\right)\\ &= 0. \end{aligned} \tag{5.25}$$

Therefore, when the number of users M is very large, we have

$$\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} \approx a_M = B\log_2\left(1+\beta F_\Gamma^{-1}(1-\frac{1}{M})\right), \tag{5.26}$$

which is a rough estimation for the average total throughput with a large M. According to (5.24), we have

$$F_\Gamma^{-1}(1-\frac{1}{M}) = \mathbb{E}\{\Gamma_{\text{eff}}\} + o(\mathbb{E}\{\Gamma_{\text{eff}}\}).$$

Thus, (5.26) can also be rewritten as

$$\begin{aligned}\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} &\approx B\log_2\left(1+\beta\left[\mathbb{E}\{\Gamma_{\text{eff}}\}+o(\mathbb{E}\{\Gamma_{\text{eff}}\})\right]\right),\\ &\approx B\log_2\left(1+\beta\mathbb{E}\{\Gamma_{\text{eff}}\}\right). \end{aligned} \tag{5.27}$$

The above equation means that the average throughput is approximately a function of the average effective SNR.

Lemma 5.2 also shows that for any positive real number p,

$$\lim_{M\to\infty}\mathbb{E}\left(\frac{R_{\text{total}}^{\text{hom}}-a_M}{b_M}\right)^p = \mathbb{E}\{Z^p\}, \tag{5.28}$$

where Z is a normalized Gumbel random variable. We consider $p=2$, and we have

$$\lim_{M\to\infty}\mathbb{E}\left(\frac{R_{\text{total}}^{\text{hom}}-a_M}{b_M}\right)^2 = E_0^2+\frac{\pi^2}{6}.$$

Thus, as $M\to\infty$,

$$\mathbb{V}ar\{R_{\text{total}}^{\text{hom}}\} \to \frac{\pi^2}{6}b_M^2.$$

Because of (5.25), $\mathbb{V}ar\{R_{\text{total}}^{\text{hom}}\} \to 0$, which indicates that this asymptotic analysis of average throughput is quite accurate. In addition, it follows that

$$\lim_{M\to\infty} \mathbb{E}\left(R_{\text{total}}^{\text{hom}} - a_M\right)^p = \lim_{M\to\infty} b_M \mathbb{E}\{Z^p\}, \tag{5.29}$$

$$= 0. \tag{5.30}$$

According to [51], (5.29) guarantees that $R_{\text{total}}^{\text{hom}} - a_M$ converges in probability[1] to 0.

5.2.3.3 Channel Access Probability and Average Throughput per User

The channel access probability P_i is the probability that user i obtains the channel to transmit data. In the homogeneous fading case, due to the the symmetry, each user has the same channel access probability; that is,

$$P_i = \frac{1}{M}.$$

Therefore, the average throughput of user i with the scheduling, $\mathbb{E}\{R_i^{\text{s}}\}$, is given by

$$\mathbb{E}\{R_i^{\text{s}}\} = \frac{1}{M}\mathbb{E}\{R_{\text{total}}^{\text{hom}}\}.$$

5.2.4 Throughput Analysis for Normalized-SNR-Based Scheduling

In previous sections, we presented the asymptotic throughput analysis for the homogeneous fading case. In reality, the values of the average SNR of users vary according to their path losses. Denote the average SNR of user i as γ_i. We consider a scenario in which different users have the same normalized SNR distribution $F(\gamma)$ but with different average SNR, γ_i's. We assume that $F(\gamma)$ satisfies $\omega(F) = \infty$ and (5.17).

Obviously, the MSC scheduling results in unfair channel access probabilities. When the normalized-SNR-based scheduling is used, the base station schedules the user with the largest normalized SNR to get the channel, which is mathematically expressed in (5.5). Define the effective normalized SNR at the transmitter as

$$\Gamma_{\text{eff}} = \max_{i\in\mathcal{M}} \frac{\Gamma_i}{\gamma_i}.$$

[1] Assume X_n and X to be a random variable sequence and a random variable, if $\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \epsilon\} = 0$ for any $\epsilon > 0$, then we say that X_n converges in probability to X.

Because of the identical distribution of the normalized SNR, the previous results based on extreme value theory is still applicable to the effective normalized SNR, and all users have the same channel access probability as well; that is,

$$P_i = \frac{1}{M}.$$

Thus, the average throughput of user i can be expressed as

$$\mathbb{E}\{R_i^{\mathrm{s}}\} = \frac{1}{M}\int_0^\infty B\log_2(1+\beta\gamma_i\gamma) f_{\Gamma_{\mathrm{eff}}}(\gamma)d\gamma \tag{5.31}$$

Recalling (5.10) and the LTD theorem, we know that in the i.i.d. fading case if the distribution of SNR $F_\Gamma(\gamma)$ satisfies (5.17), then, with a large M,

$$\int_0^\infty \log_2(1+\beta\gamma) f_{\Gamma_{\mathrm{eff}}}(\gamma)d\gamma \approx \log_2\left(1+\beta F_\Gamma^{-1}(1-\frac{1}{M})\right) + E_0\log_2\left(\frac{1+\beta F_\Gamma^{-1}(1-\frac{1}{Me})}{1+\beta F_\Gamma^{-1}(1-\frac{1}{M})}\right). \tag{5.32}$$

Comparing (5.31) and (5.32), we obtain the average throughput for user i as follows:

$$\mathbb{E}\{R_i^{\mathrm{s}}\} \approx \frac{B}{M}\left\{\log_2\left(1+\beta\gamma_i F^{-1}(1-\frac{1}{M})\right) + E_0\log_2\left(\frac{1+\beta\gamma_i F^{-1}(1-\frac{1}{Me})}{1+\beta\gamma_i F^{-1}(1-\frac{1}{M})}\right)\right\},$$

with a large M. Therefore, with the normalized-SNR-based scheduling, each user obtains the same multiuser diversity gain as that in the homogeneous scenario and has the same channel access probability, but its own average throughput depends on its average SNR.

Furthermore, we will compare the total throughput in the heterogeneous and homogeneous scenarios. We assume that

$$\gamma_0 = \frac{1}{M}\sum_{i=1}^M \gamma_i, \tag{5.33}$$

and define

$$\sigma_\gamma^2 = \frac{1}{M}\sum_{i=1}^M (\gamma_i - \gamma_0)^2.$$

When the number of users M is large, we only consider use the first term, a_M, to evaluate the average throughput. Thus, the average total throughput in the heterogeneous scenario is given by

$$\begin{aligned}\mathbb{E}\{R_{\mathrm{total}}^{\mathrm{het}}\} &= \sum_{i=1}^M E\{R_i^{\mathrm{s}}\} \\ &\approx \frac{B}{M}\sum_{i=1}^M \log_2\left(1+\beta\gamma_i F^{-1}(1-\frac{1}{M})\right),\end{aligned}$$

and the average total throughput in the homogeneous scenario is

$$\mathbb{E}\{R_{\text{total}}^{\text{hom}}\} \approx B \log_2 \left(1 + \beta\gamma_0 F^{-1}(1 - \frac{1}{M})\right),$$

We obtain

$$\begin{aligned}\mathbb{E}\{R_{\text{total}}^{\text{het}}\} - \mathbb{E}\{R_{\text{total}}^{\text{hom}}\} &\approx \frac{B}{M}\sum_{i=1}^{M} \log_2 \left(\frac{1 + \beta\gamma_i F^{-1}(1 - \frac{1}{M})}{1 + \beta\gamma_0 F^{-1}(1 - \frac{1}{M})}\right) \\ &\to \frac{B}{M}\sum_{i=1}^{M} \log_2 \left(\frac{\gamma_i}{\gamma_0}\right), \quad \text{as } M \to \infty. \end{aligned} \tag{5.34}$$

(5.34) is valid since $F^{-1}(1 - \frac{1}{M}) \to \infty$ as $M \to \infty$.

With the following inequality,

$$x - \frac{1}{2}x^2 \leq \ln(1 + x) \leq x, \quad \text{for } x \geq 0, \tag{5.35}$$

we will consider the upper and lower bounds, respectively. For the upper bound, it follows from (5.34) and (5.35) that

$$\begin{aligned}\mathbb{E}\{R_{\text{total}}^{\text{het}}\} - \mathbb{E}\{R_{\text{total}}^{\text{hom}}\} &\leq \frac{B}{\ln(2)M}\sum_{i=1}^{M}\left(\frac{\gamma_i}{\gamma_0} - 1\right) \\ &= 0.\end{aligned}$$

Similarly, the lower bound is given by

$$\begin{aligned}\mathbb{E}\{R_{\text{total}}^{\text{het}}\} - \mathbb{E}\{R_{\text{total}}^{\text{hom}}\} &> \frac{B}{\ln(2)M}\sum_{i=1}^{M}\left(\frac{\gamma_i}{\gamma_0} - 1\right) - \frac{1}{2\ln 2}\frac{B}{M}\sum_{i=1}^{M}\left(\frac{\gamma_i}{\gamma_0} - 1\right)^2 \\ &= 0 - \frac{B}{2\ln 2}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\gamma_i}{\gamma_0}\right)^2 - 1\right] \\ &= -\frac{B}{2\ln 2}\frac{\sigma_\gamma^2}{\gamma_0^2}.\end{aligned}$$

Therefore, the main result is stated as follows: when the number of users M is large,

$$-\frac{B}{2\ln 2}\frac{\sigma_\gamma^2}{\gamma_0^2} \leq \mathbb{E}\{R_{\text{total}}^{\text{het}}\} - \mathbb{E}\{R_{\text{total}}^{\text{hom}}\} \leq 0. \tag{5.36}$$

This means that the homogeneous case leads to the maximum total throughput when (5.33) holds.

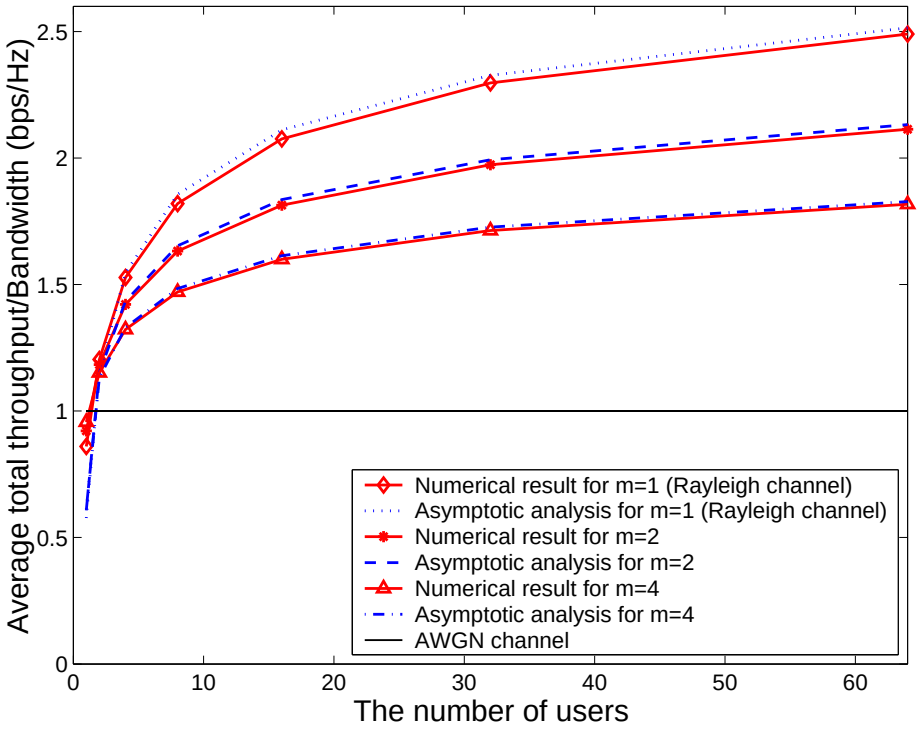


Figure 5.1. Average throughput for different environments. $\beta\gamma_0 = 1$.

5.2.5 Numerical Results

We assume that all users experience i.i.d. Nakagami fading. Let $\beta\gamma_0 = 1$. Figure 5.1 shows the average total throughput in the Nakagami fading channels with different values of m. For comparison, we also plot the average throughput in the *additive white Gaussian noise* (AWGN) channel with the same average SNR in Figure 5.1.

It is shown in Figure 5.1 that the asymptotic results are still accurate even if the number of users is small. The figure shows that the throughput increases with the number of users in the fading scenario with dynamic scheduling. As m increases, the fading fluctuation of the channel reduces, and the multiuser diversity gain is also diminished.

5.3 Asymptotic Delay Analysis of Single-Carrier Networks

Besides throughput, delay is another crucial factor for wireless data services, particularly for time-sensitive applications. In [39], an analysis for mean delays is presented; however, only a rough relationship between the average waiting time and the number of users in the system

is provided. In [11], the system performance in the mean sense is studied by using multi-class processor-sharing model. Since the dynamic user configuration is considered in [11], the results are applicable for more general cases, but the complicated model has difficulty in capturing the explicit relationship between system performance and multiuser diversity. In this section, we propose an asymptotically analytical result to reveal the impact of dynamic scheduling on average waiting times in single-carrier networks. Compared to [11], our analysis has two major differences. First, we just consider the static user scenario to reveal the impact of multiuser diversity on the mean delay. Although our analysis overestimates the mean delay with a light traffic load, they would be accurate with a heavy traffic load. Second, the system performance is evaluated in terms of the average delay for each packet rather than the average delay for each file transmission in [11].

The system and scheduling models are the same as those in Section 5.2.1. In our analysis, we make the following assumptions.

- When a queue is empty, a dummy packet is assumed to be in the queue. This system is usually called the *dominant system*.
- Dynamic packet scheduling is usually allowed in time-slotted networks, which makes delay analysis extremely difficult. For simplicity, we assume that the system is not time-slotted. After finishing transmitting a packet, the base station can immediately serve another packet.
- We assume that all users have the same channel statistics and arrival traffic statistics.
- For the arrival traffic, all packets are assumed to have the same length L. In addition, the packet inter-arrival time for each user is assumed to be independently, identically, and exponentially distributed with rate λ_1. The total arrival rate is $\lambda = M\lambda_1$.
- For the channel fading, we assume that each user experiences i.i.d. block fading that is constant while a packet is being served, but is independent across different packet transmission durations.

5.3.1 Asymptotic Distribution of Service Time

The service time S for transmitting a packet can be expressed as

$$\begin{aligned} S &= S(\Gamma) \\ &= \frac{L}{R} \\ &= \frac{L}{B}\frac{1}{\log_2(1+\beta\Gamma)}. \end{aligned}$$

Thus, the distribution of the service time is

$$F_S(s) = 1 - F_\Gamma(S^{-1}(s)),$$

and its inverse function is

$$\begin{aligned} F_S^{-1}(x) &= S(F_\Gamma^{-1}(1-x)) \\ &= \frac{L}{B}\frac{1}{\log_2(1+\beta F_\Gamma^{-1}(1-x))}. \end{aligned}$$

According to the MSC rule in the multiuser-symmetric environment, the base station should serve the user with the strongest channel condition. This is equivalent to serving the user who needs the shortest service time. Since the system performance in the scenario of a large number of users can be obtained by extreme value theory [13], we focus on the properties of the limit distribution of the random variable

$$S_{min,M} = \min_{i\in\mathcal{M}} S_i,$$

where S_i is the service time for user i.

Similar to the LTD theorem, we state the result about asymptotic distribution of service time as the following theorem.

Theorem 5.2 *Assume that all users' SNRs, $\{\Gamma_1, \Gamma_2, \cdots, \Gamma_M\}$, are i.i.d. random variables with a distribution $F_\Gamma(\gamma)$ such that $\omega(F_\Gamma) = \infty$, and $f_\Gamma(\gamma) = F'_\Gamma(\gamma)$ as well as $F''_\Gamma(\gamma)$ exist and $f_\Gamma(\gamma) \neq 0$ for all $x_1 \leq x < \infty$, where x_1 is some real number. If*

$$\lim_{\gamma\to\infty} \frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right] = 0, \tag{5.37}$$

then the distribution of service, $F_S(s) = 1 - F_\Gamma(S^{-1}(s))$, belongs to the domain of the attraction of the Gumbel distribution (minima). In addition,

$$c_M = \frac{L}{B}\frac{1}{\log_2(1+\beta F_\Gamma^{-1}(1-\frac{1}{M}))}, \tag{5.38}$$

$$d_M = \frac{L}{B}\frac{1}{\log_2(1+\beta F_\Gamma^{-1}(1-\frac{1}{M}))} - \frac{L}{B}\frac{1}{\log_2(1+\beta F_\Gamma^{-1}(1-\frac{1}{Me}))}. \tag{5.39}$$

$$\lim_{M\to\infty}\frac{d_M}{c_M} = 0. \tag{5.40}$$

The proof is omitted since it is very similar to that of LTD theorem. The fact that $c_M \to 0$ as $M \to \infty$, together with (5.40), implies that $S_{min,M}$ is approximately a constant when M is large enough; that is,

$$S_{min,M} \approx c_M - E_0 d_M.$$

In the case of Rayleigh fading, we have

$$c_M = \frac{L}{B}\frac{1}{\log_2(1+\beta\gamma_0\ln(M))},$$
$$d_M = \frac{L}{B}\frac{1}{\log_2(1+\beta\gamma_0\ln(M))} - \frac{L}{B}\frac{1}{\log_2\left(1+\beta\gamma_0\left(1+\ln(M)\right)\right)}.$$

Therefore, we obtain the service rate for one packet.

5.3.2 Average Waiting Time

Because the stochastic characteristics of the packet arrivals and wireless channels of different users are symmetrical, we only need to consider the delay performance of a specific user. Under the MSC rule, each queue (user) is equally served with probability $\frac{1}{M}$. As a result, the time needed to transmit one packet, X, which is an integer multiple of $S_{min,M}$, has a geometric distribution,

$$\mathbb{P}(X = nS_{min,M}) = \left(\frac{1}{M}\right)\left(1-\frac{1}{M}\right)^{n-1},$$

where n is an integer.

Due to the Poisson arrivals, each queue can be modeled as an M/G/1 queue with server vacations [10]. In equilibrium, the mean waiting time in a queue, W_q, can be decomposed into the expected residual service time T_{res} plus the average service time of packets in the

queue $\mathbb{E}\{X\}N_q$, where N_q is the average queue length. Applying Little's law, $N_q = \lambda_1 W_q$, we have

$$\begin{aligned} W_q &= T_{res} + \mathbb{E}\{X\}\lambda_1 W_q \\ &= T_{res} + \rho W_q, \end{aligned}$$

where $\rho = \lambda S_{min,M} = M\lambda_1 S_{min,M}$. Therefore, the mean waiting time in queue is given by

$$W_q = \frac{T_{res}}{1-\rho}. \tag{5.41}$$

Using the results of M/G/1 queues with vacations in [10], the expected residual service time T_{res} can obtained as

$$T_{res} = \frac{\lambda_1 \mathbb{E}\{X^2\}}{2} + (1-\rho)\frac{\mathbb{E}\{V^2\}}{2\mathbb{E}\{V\}},$$

where the length of a server vacation, V, is equal to $S_{min,M}$. Thus, it follows that

$$\begin{aligned} T_{res} &= \frac{\lambda_1 (2M-1) M S_{min,M}^2}{2} + (1-\rho)\frac{S_{min,M}}{2} \\ &= \frac{(2M-1)\rho S_{min,M}}{2} + (1-\rho)\frac{S_{min,M}}{2}. \end{aligned}$$

The mean waiting time in the single-carrier system, W_{single}, includes the mean waiting time in queue and the service time of transmitting one packet. Thus,

$$\begin{aligned} W_{\text{single}} &= \frac{T_{res}}{1-\rho} + \mathbb{E}\{X\} \\ &= \frac{(2M-1)\rho S_{min,M}}{2(1-\rho)} + \left(M + \frac{1}{2}\right) S_{min,M}. \end{aligned} \tag{5.42}$$

5.4 Asymptotic Performance Analysis of Multicarrier Networks

In a multicarrier network with the same bandwidth B, the scheme assigns each subchannel to the user with the best channel condition on it, which can be expressed as

$$m(k) = \arg\max_{i \in \mathcal{M}} \{\Gamma_i[k]\},$$

where $m(k)$ represents the user scheduled at subcarrier k, and $\Gamma_i[k]$ is the SNR of user i at subcarrier k. Let $F_\Gamma(\gamma)$ be the distribution of the channel fading at each subcarrier. Other assumptions here are the same as in the single-carrier network in Sections 5.2 and 5.3.

5.4.1 Asymptotic Throughput Analysis

Note that there is no assumption on the correlation among subcarriers. The data rate at subcarrier k is given by

$$R_{\max}[k] = \max_{i \in \mathcal{M}} \frac{B}{K} \log_2(1 + \beta \Gamma_i[k]). \tag{5.43}$$

Then, the total throughput is given by

$$R_{\text{total}} = \sum_{k=1}^{K} R_{\max}[k].$$

Thus,

$$\begin{aligned} \mathbb{E}\{R_{\text{total}}\} &= K\mathbb{E}\{R_{\max}[k]\} \\ &= \mathbb{E}\{\max_{i \in \mathcal{M}} B \log_2(1 + \beta\Gamma)\}, \end{aligned}$$

where Γ is distributed with $F_\Gamma(\gamma)$. It follows from Theorem 5.1 that with a large M,

$$\mathbb{E}\{R_{\text{total}}\} \approx a_M + E_0 b_M, \tag{5.44}$$

where a_M and b_M are determined by (5.18) and (5.19). Therefore, the multicarrier network has the same asymptotic throughput as the single-carrier network with the same bandwidth.

5.4.2 Asymptotic Delay Analysis

Similarly in the single-carrier system, each user has a probability $\frac{1}{M}$ of occupying a subcarrier. We consider an "extreme" scenario where the channel fluctuations are independent across the subcarriers, and the number of subcarriers $K \to \infty$. At subcarrier k, the resulting data rate is a random variable given by (5.43). According to the strong law of large numbers, as $K \to \infty$ and $M \ll K$, the total throughput is obtained as

$$\begin{aligned} R_{\text{total}} &= \frac{B}{K} \sum_{k=1}^{K} \max_{i \in \mathcal{M}} \log_2(1 + \beta \Gamma_i[k]) \\ &\to B\, \mathbb{E}\{\max_{i \in \mathcal{M}} \log_2(1 + \beta\Gamma)\}, \text{ as } K \to \infty, \end{aligned}$$

where Γ is distributed with $F_\Gamma(\gamma)$. Therefore,

$$R_{\text{total}} \approx a_M + E_0 b_M,$$

and the service time is

$$S'_{min,M} = \frac{L}{R_{\text{total}}}.$$

Since each user occupies a bandwidth of B/M, the MSC scheduling results in a traditional FDM system with the fixed service rate R/M for each user. In other words, the service time for one packet is $MS'_{min,M}$. Based on the results in [10], the mean waiting time in queue can be expressed as

$$\begin{aligned} W_q &= \frac{\lambda_1 M^2 S'^2_{min,M}}{2(1-\rho)} + \frac{S'_{min,M}}{2} \\ &= \frac{M\rho S'_{min,M}}{2(1-\rho)} + \frac{S'_{min,M}}{2}. \end{aligned}$$

Therefore, the average waiting time in the multicarrier network, W_{multi}, is given by

$$W_{\text{multi}} = \frac{M\rho S'_{min,M}}{2(1-\rho)} + (M + \frac{1}{2})S'_{min,M}. \tag{5.45}$$

The structure of the average waiting time expression for multicarrier networks in (5.45) is similar to that for single-carrier networks in (5.42). We will compare them in the next subsection.

5.4.3 Delay Performance Comparison

As the number of users $M \to \infty$,

$$\begin{aligned} S_{min,M} &\to c_M \\ &= \frac{L}{B} \frac{1}{\log_2(1 + \beta F_\Gamma^{-1}(1 - \frac{1}{M}))}, \\ S'_{min,M} &\to \frac{1}{a_M} \\ &= \frac{L}{B} \frac{1}{\log_2(1 + \beta F_\Gamma^{-1}(1 - \frac{1}{M}))}. \end{aligned}$$

In other words, both single and multiple carrier networks have the same asymptotic throughput

$$B \log_2(1 + \beta F_\Gamma^{-1}(1 - \frac{1}{M})).$$

However, each system has a different delay performance for bursty traffic. When the traffic load is light,

$$\lim_{\rho \to 0} \left(\lim_{\substack{M \to \infty \\ \text{fixing } \rho}} \frac{W_{\text{single}}}{W_{\text{multi}}} \right) = 1.$$

When the traffic load is heavy,

$$\lim_{\rho \to 1} \left(\lim_{\substack{M \to \infty \\ \text{fixing } \rho}} \frac{W_{\text{single}}}{W_{\text{multi}}} \right) = 2. \tag{5.46}$$

This is because multicarrier networks can provide smoother service rates by exploiting frequency diversity.

In the above analysis, we do not take the time correlation in channel fading into account. This issue will be discussed in a descriptive manner as follows. We consider the extreme case where the channel correlation time goes to infinity. In a single-carrier system using the MSC rule, the average waiting time will become infinite with a heavy traffic load. However, note that in the delay performance analysis of multicarrier networks, we exploit the channel independence among subcarriers instead of the channel independence across different packet transmission durations. Thus, for a multicarrier network in a highly frequency-selective environment, the channel time correlation does not affect the average waiting time, and the average waiting time is always equal to (5.45). Therefore, when the channel fading is slow and highly frequency-selective, the multicarrier network *greatly* outperforms the single-carrier network in terms of delay performance. More accurately, if the number of subcarriers in the multicarrier network is large, the average waiting time in the multicarrier network is half that in the single-carrier network when the traffic load is heavy, or is considerably less than half with slow fading.

Figure 5.2 shows that the average waiting time of different systems in the Rayleigh fading environment with the same bandwidth when there are 100 users. Since the MSC scheduling can improve the throughput through multiuser diversity, both single-carrier and multicarrier networks with this scheduling provide a substantial delay performance improvement, compared to the traditional TDMA. In the simulation, we consider a simple time correlation model of the fading as follows. The channel is block-faded; the channel remains constant

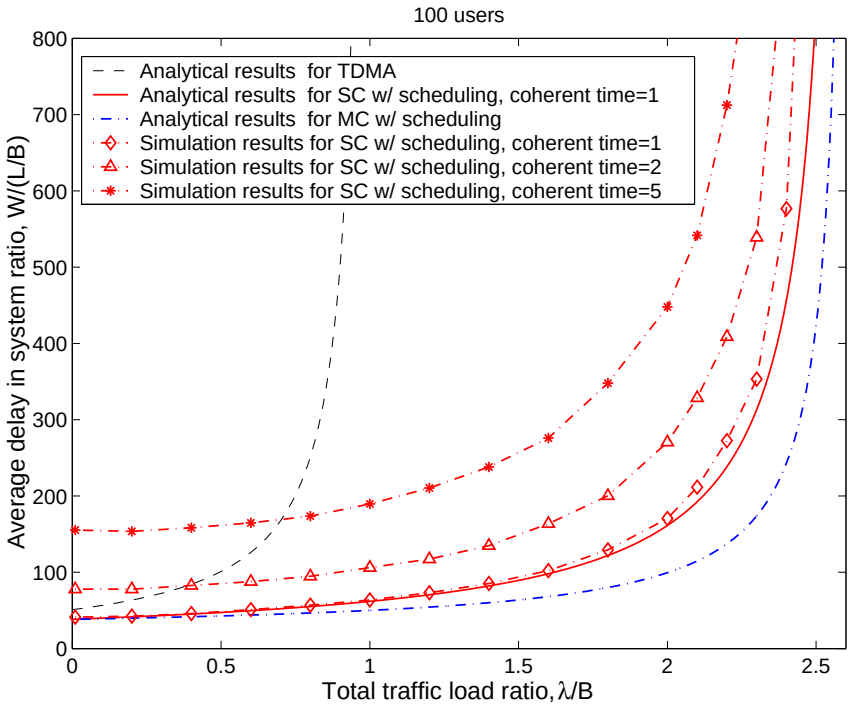


Figure 5.2. Average waiting time versus traffic load. $\beta\gamma_0 = 1$**, and** $M = 100$

within a block, but is independent across different blocks. The length of a block is the coherence time of the channel, which is an integer multiple of $S_{min,M}$. A longer coherence time indicates a slow fading rate. When the coherence time equals 1, the fading model is the same as that assumed in Section 5.2. It is concluded from Figure 5.2 that the analytical result (5.42) fits the simulation curve very well, and that slow fading seriously impairs the delay performance of single-carrier networks.

5.5 Summary

Using extreme value theory, we have proposed asymptotic average throughput and delay analyses for the MSC scheduling with a general fading distribution in both single-carrier and multicarrier networks, which not only have concise expressions, but also provide accurate results. This asymptotic analysis shows that the use of the simple scheduling techniques and the feedback of CSI can significantly improve the bandwidth efficiency. We have also extended the analysis into a scenario in which different users experience different path losses. The results shows that the normalized-SNR-based scheduling can obtain the same

multiuser diversity gain as that in the homogeneous case while maintaining access-time proportional fairness. Although multicarrier networks with channel-aware scheduling have the same throughput as single-carrier networks, multicarrier networks can provide better delay performance than single-carrier networks. This work is beneficial to QoS provisioning for channel-aware scheduling.

CHAPTER 6

CONCLUSION

This final chapter summarizes the major contributions of the thesis and highlights numerous topics for future research.

6.1 Contributions

In this thesis, we have investigated resource allocation and scheduling in the wireless OFDM-based downlink that serves multiple users and supports various applications based on joint physical and MAC layer optimization. The main contributions of this thesis are summarized as follows:

We have proposed a simple, effective cross-layer resource management architecture for wireless resource allocation. The use of rate adaptation and packet scheduling can exploit the characteristics of wireless channels, such as time variance, frequency selectivity, and statistical independence among different users, to obtain the available natural diversity - multiuser diversity. More importantly, with the help of economic theory, we use utility as a measure of QoS and maximize the total utility in the network based on the current channel and QoS conditions. Both theoretical and simulation results show that the utility-based architecture can provide an efficient and stable mechanism for spectral efficiency improvement, traffic multiplexing, and QoS differentiation.

We have developed various efficient DSA and APA algorithms that solve the proposed utility optimization problems in multicarrier networks with different considerations, especially for the scenario with discrete rate adaptation, in which the nonlinear and combinatorial nature of the cross-layer optimization significantly challenges algorithm development. Moreover, those algorithms are low-complexity and stable.

We have designed a novel scheduling approach maximizing the total utility with respect

to mean delays, which is call MDU scheduling. Unlike most joint channel- and queue-aware scheduling policies, such as the M-LWDF and EXP rules, the MDU scheduling has an explicit optimization objective. Although it does not need statistical information about incoming traffic, its utility-maximization mechanism enables the network to achieve the right balance between capacity enlargement and QoS guarantees according to the channel conditions and the level of network congestion. In Chapter 4, the MDU scheduling was successfully used for handling heterogeneous traffic. The simulation results demonstrated that the MDU scheduling can improve the spectral efficiency and provide right incentives to ensure that all applications can receive their different required QoS. Through the thread of cross-layer design, we also proposed delay transmit diversity, a simple and transparent multiple transmit antenna technique that enhances the performance of the multicarrier scheduling without requirement on algorithm modification.

We have deeply studied the fairness and stability issues in a general sense. First, we have in theory revealed a generic relationship between a specific convex utility function and a type of fairness. It is shown that the utility-based resource allocation has an explicit fairness characteristic. Second, we proved that given some very loose conditions, the maximum stability region of stable scheduling policies can reach the interior of the ergodic capacity region at the physical layer. The results and proofs concerned with the stability issue in this thesis are applicable to both cases of single and multiple servers and do not require the Markovian property on channel states and/or arrival traffic. More importantly, we provided a method to design cross-layer scheduling algorithms that allow the queueing stability region at the network layer to approach the ergodic capacity region at the physical layer. In Chapter 4, the results of the stability issue was successfully used in designing the MDU scheduling for supporting integrated services.

To reveal the impact of multiuser diversity on throughput and delay performance, we have provided closed-form asymptotic performance analyses for channel-aware scheduling in both single-carrier and multicarrier networks based on extreme value theory and queueing theory. Compared to the exact expression, the asymptotic one, which is applicable to a broader range of fading channels, is more concise and easier to get insights. One of the

most important results is that, in environments with highly frequency-selective fading, slow fading does not limit the multiuser diversity gain in multicarrier networks with DSA.

In summary, this thesis is a deep and systematic study on cross-layer resource allocation and scheduling in wireless OFDM-based networks. It not only proposes a utility-based cross-layer wireless resource management architecture and corresponding scheduling algorithms that substantially improve the spectral efficiency and effectively satisfy diverse performance objectives of heterogeneous traffic, but also provides deep understanding of fundamental mechanisms in advanced wireless resource management, including efficiency, fairness, and stability, which would facilitate the design of future wireless multimedia networks that support diverse QoS requirements in such a complicated environment where multiple users compete shared channels with time-varying frequency-selective fading.

6.2 Future Research Directions

Cross-layer resource allocation is promising for future wireless networks. Two important open questions are list as follows.

6.2.1 Admission Control for Channel-Aware Scheduling and MAC

The mechanism of exploiting channel variations across users has been used in scheduling and MAC designs to improve the spectral efficiency. Due to variable data rates and stochastic transmission inherent in channel-aware networks, the issue of admission control is becoming very challenging and interesting. Our work on efficiency, fairness, and stability of channel-aware scheduling will be beneficial to studying this issue. This research will result in theoretical innovations and practical applications because this topic may lead to rethinking the architecture of multimedia-over-wireless networks, and because current CDMA2000 1xEV systems require commercial solutions for admission control.

6.2.2 Distributed Channel- and QoS-Aware Multicarrier MAC

Since the resource allocation schemes studied in this thesis require centralized control, it would be of great interest to extend our research to developing scalable and distributed channel- and QoS-aware multicarrier MAC schemes without a centralized controller. The

research should be based on deep understanding of specific properties of multicarrier systems. Distributed channel- and QoS-aware MAC approaches are very promising for the following two major reasons. First, channel-aware-only MAC schemes are only optimal for the total throughput rather than diverse QoS requirements of different applications. On the other hand, the queueing model of a single carrier system is a single server with a queue. However, since multicarrier systems can serve many users at the same time, there are multiple servers from a queueing theory point of view. The multiserver systems would be advantageous to QoS provisioning.

APPENDIX A

PROOF OF THEOREM 2.1

Proof: If the $\bar{D}_1^*$'s are optimal, then any change of allocation will not increase the average utility. Let $(f - \frac{1}{2}\triangle f,\ f + \frac{1}{2}\triangle f) \in D_1^*$. If $(f - \frac{1}{2}\triangle f,\ f + \frac{1}{2}\triangle f)$ is assigned to the other user, then the data rate of user 1 will be decreased by $\triangle r_1 = c_1(f)\triangle f$ while the data rate of user 2 will be increased by $\triangle r_2 = c_2(f)\triangle f$. But, the new average utility will be equal to or less than the optimal one, that is,

$$U_1(r_1^* - \triangle r_1) + U_2(r_2^* + \triangle r_2) \leq U_1(r_1^*) + U_2(r_2^*),$$

which is equivalent to

$$U_2(r_2^* + \triangle r_2) - U_2(r_2^*) \leq U_1(r_1^*) - U_1(r_1^* - \triangle r_1).$$

Dividing both sides by $\triangle f$, we have

$$\frac{U_2(r_2^* + \triangle r_2) - U_2(r_2^*)}{\triangle f} \leq \frac{U_1(r_1^*) - U_1(r_1^* - \triangle r_1)}{\triangle f}.$$

Since $\triangle r_1 = c_1(f)\triangle f$ and $\triangle r_2 = c_2(f)\triangle f$, we have

$$\frac{U_2(r_2^* + \triangle r_2) - U_j(r_2^*)}{\triangle r_2} c_2(f) \leq \frac{U_1(r_1^*) - U_1(r_1^* - \triangle r_1)}{\triangle r_1} c_1(f).$$

When $\triangle f \to 0$, $\triangle r_1 \to 0$ and $\triangle r_2 \to 0$. Consequently,

$$\lim_{\triangle r_2 \to 0} \frac{U_2(r_2^* + \triangle r_2) - U_j(r_2^*)}{\triangle r_2} c_2(f) \leq \lim_{\triangle r_1 \to 0} \frac{U_1(r_1^*) - U_1(r_1^* - \triangle r_1)}{\triangle r_1} c_1(f),$$

or

$$U_2^{'}(r_2^*)c_2(f) \leq U_1^{'}(r_1^*)c_1(f) \quad f \in \bar{D}_1^*. \tag{A.1}$$

which implies, for any $f \in \bar{D}_1^*$,

$$\frac{c_2(f)}{c_1(f)} \leq \frac{U_1^{'}(r_1^*)}{U_2^{'}(r_2^*)} (= \alpha^*),$$

that is, $f \in \bar{D}_1(\alpha^*)$ and $D_1^* \subseteq \bar{D}_1(\alpha^*)$.

Similarly, we can prove that

$$D_2^* \subseteq \bar{D}_2(\alpha^*).$$

Therefore,

$$\begin{aligned} D_1(\alpha^*) &= [0,\ B] - \bar{D}_2(\alpha^*) \\ &\subseteq [0,\ B] - D_2^* \\ &= D_1^* \end{aligned}$$

□

APPENDIX B

PROOF OF THEOREM 2.3

Proof: For a fixed subcarrier assignment D_i for all i, we define $p_i(f)$ for $i = 1,\ 2, \cdots,\ M$ as,

$$p_i(f) = \begin{cases} p(f) & f \in D_i \\ 0 & \text{otherwise} \end{cases}.$$

Using the Lagrangian method, the above optimization problem with the power constraint becomes to maximize

$$\frac{1}{M}\sum_{i=1}^{M} U_i\left(\int_{D_i}\log_2\left[1+\beta p(f)\rho_i(f)\,df\right]\right) - \lambda^{'}\left[\frac{1}{B}\int_0^B p(f)\,df - 1\right],$$

or

$$\frac{1}{M}\sum_{i=1}^{M}\left\{U_i\left(\int_{D_i}\log_2\left[1+\beta p_i(f)\rho_i(f)\,df\right]\right) - \lambda^{'}\left[\frac{1}{B}\int_{D_i} p_i(f)\,df - 1\right]\right\}.$$

where $\lambda^{'} \geq 0$.

With the Karush-Kuhn-Tucker (KKT) conditions [56], we have

$$\frac{1}{M}U_i^{'}(r_i^*)\frac{\partial}{\partial p_i(f)}\log_2\left\{1+\beta p_i(f)\rho_i(f)\right\} - \frac{\lambda^{'}}{B}\frac{\partial}{\partial p_i(f)}p_i(f)\Big|_{p_i(f)=p_i^*(f)} = 0, \quad \text{for all } i, \tag{B.1}$$

$$\lambda^{'} \geq 0, \tag{B.2}$$

$$\lambda^{'}[\sum_{i=1}^{M}\frac{1}{B}\int_{D_i} p_i(f)\,df - 1] = 0. \tag{B.3}$$

(B.1) is equivalent to

$$U_i^{'}(r_i^*)\frac{\beta\rho_i(f)}{1+\beta\rho_i(f)p_i^*(f)} - \lambda^{'}\frac{M}{\log_2(e)B} = 0, \quad \text{for all } i.$$

Let $\lambda = \lambda^{'}\frac{M}{\log_2(e)B}$. Then, the optimal power allocation for a fixed subcarrier assignment satisfies:

$$\begin{cases} p_i^*(f) = \left[\dfrac{U_i^{'}(r_i^*)}{\lambda} - \dfrac{1}{\beta\rho_i(f)}\right]^+ & f \in D_i \\ \displaystyle\sum_{i=1}^{M}\frac{1}{B}\int_{D_i} p_i^*(f)\,df = 1. \end{cases}$$

or

$$\begin{cases} p^*(f) = \left[\dfrac{U_i'(r_i^*)}{\lambda} - \dfrac{1}{\beta\rho_i(f)}\right]^+ & f \in D_i \\ \dfrac{1}{B}\displaystyle\int_0^B p^*(f)\,df = 1. \end{cases}$$

□

APPENDIX C

PROOF OF THEOREM 2.5

Proof: Assume that the system has joint DSA and APA. Then $\forall\ \mathbf{r}^{(1)},\ \mathbf{r}^{(2)} \in \mathcal{C}_{DSA+APA}$, $\alpha \in [0,1]$, we need to show that $\alpha\mathbf{r}^{(1)} + (1-\alpha)\mathbf{r}^{(2)} \in \mathcal{C}_{DSA+APA}$. $\mathbf{r}^{(1)} = [r_1^{(1)}, r_2^{(1)}, \ldots, r_M^{(1)}]^T$ is achieved with $D_m^{(1)}$ and $p^{(1)}(f)$, $\mathbf{r}^{(2)} = [r_1^{(2)}, r_2^{(2)}, \ldots, r_M^{(2)}]^T$ is achieved with $D_m^{(2)}$ and $p^{(2)}(f)$, where for $m \in \{1, 2, \cdots, M\}$. Of course, $D_m^{(1)}$ and $D_m^{(2)}$ satisfy (2.3) and (2.4); $p^{(1)}(f)$ and $p^{(2)}(f)$ yield (2.5). We represent those two power allocations as $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, respectively.

We define the measure of a frequency set as follows. When the frequency set $D = \bigcup_i [a_i, b_i], \quad b_i \le a_{i+1}$, the measure μ is given by $\mu(D) = \sum_i (b_i - a_i)$. For user m, we have

$$r_m^{(1)} = \int_{D_m^{(1)}} c_m^{\mathbf{p}^{(1)}}(f)\, d\mu$$
$$r_m^{(2)} = \int_{D_m^{(2)}} c_m^{\mathbf{p}^{(2)}}(f)\, d\mu$$

where $c_m^{\mathbf{p}}(f)$ denotes the achievable throughput of user m at frequency f with power allocation $\mathbf{p}$.

We divide $[0, B]$ into a family of sets F_n's so that

$$\bigcup_n F_n = [0, B],\ F_i \bigcap F_j = \varnothing\ i \ne j \tag{C.1}$$

$$\max_{f \in F_n} \{c_m^{\mathbf{p}^{(1)}}(f)\} - \min_{f \in F_n} \{c_m^{\mathbf{p}^{(1)}}(f)\} \to 0 \text{ for all } m, n \tag{C.2}$$
$$\max_{f \in F_n} \{c_m^{\mathbf{p}^{(2)}}(f)\} - \min_{f \in F_n} \{c_m^{\mathbf{p}^{(2)}}(f)\} \to 0 \text{ for all } m, n. \tag{C.3}$$

(C.2) and (C.3) imply

$$\max_{f \in F_n} p^{(1)}(f) - \min_{f \in F_n} p^{(1)}(f) \to 0 \text{ for all } \ n$$
$$\max_{f \in F_n} p^{(2)}(f) - \min_{f \in F_n} p^{(2)}(f) \to 0 \text{ for all } \ n.$$

Each F_n is divided into two subsets F_n^{α} and $F_n^{(1-\alpha)}$ that satisfy

$$F_n^{\alpha} \bigcup F_n^{(1-\alpha)} = F_n,\ F_n^{\alpha} \bigcap F_n^{(1-\alpha)} = \varnothing \tag{C.4}$$

and $\mu(F_n^{\alpha}) = \alpha\mu(F_n)$.

If $F_n \in D_m$, we use $D_{m,n}$ to denote F_n. Thus,

$$r_m^{(1)} = \sum_n c_m^{\mathbf{p}^{(1)}}(n)\mu(D_{m,n}^{(1)})$$
$$r_m^{(2)} = \sum_n c_m^{\mathbf{p}^{(2)}}(n)\mu(D_{m,n}^{(2)})$$

In the same way, using $D_{m,n}^{\alpha}$ to denote $F_n^{\alpha} \in D_m$, we have

$$\int_{D_m^{(1),\alpha}} c_m^{\mathbf{p}^{(1)}}(f)\, d\mu = \sum_n c_m^{\mathbf{p}^{(1)}}(n)\mu(D_{m,n}^{(1),\alpha})$$
$$= \alpha r_m^{(1)}$$

$$\int_{D_m^{(2),(1-\alpha)}} c_m^{\mathbf{p}^{(2)}}(f) d\mu = \sum_n c_m^{\mathbf{p}^{(2)}}(n)\mu(D_{m,n}^{(2),(1-\alpha)})$$
$$= (1-\alpha) r_m^{(2)}$$

$$\text{where} \qquad D_m^{(1),\alpha} = \bigcup_n D_{m,n}^{(1),\alpha}$$
$$D_m^{(2),(1-\alpha)} = \bigcup_n D_{m,n}^{(2),(1-\alpha)}.$$

Therefore, with the new frequency assignment $D_m = D_m^{(1),\alpha} \cup D_m^{(2),(1-\alpha)}$ and the new power allocation

$$p(f) = \begin{cases} p^{(1)}(f) & f \in D_m^{(1),\alpha} \\ p^{(2)}(f) & f \in D_m^{(2),(1-\alpha)} \end{cases},$$

the new data rate for user m is

$$\begin{aligned} r_m &= \int_{D_m^{(1),\alpha}} c_m^{\mathbf{p}^{(1)}}(f)\, d\mu + \int_{D_m^{(2),(1-\alpha)}} c_m^{\mathbf{p}^{(2)}}(f)\, d\mu \\ &= \alpha r_m^{(1)} + (1-\alpha) r_m^{(2)} \end{aligned}$$

Furthermore, due to (C.1) and (C.4), the D_m's satisfy (2.3) and (2.4). In addition,

$$\begin{aligned}
& \frac{1}{B}\int_0^B p(f)\,d\mu = \frac{1}{B}\sum_m \int_{D_m} p(f)\,d\mu \\
= \;& \frac{1}{B}\sum_m \int_{D_m^{(1),\alpha}} p^{(1)}(f)\,d\mu + \frac{1}{B}\sum_m \int_{D_m^{(2),(1-\alpha)}} p^{(2)}(f)\,d\mu \\
= \;& \frac{\alpha}{B}\sum_m \int_{D_m^{(1)}} p^{(1)}(f)d\mu + \frac{1-\alpha}{B}\sum_m \int_{D_m^{(2)}} p^{(2)}(f)d\mu \\
= \;& \alpha\frac{1}{B}\int_0^B p^{(1)}(f)d\mu + (1-\alpha)\frac{1}{B}\int_0^B p^{(2)}(f)\,d\mu \\
\leq \;& 1
\end{aligned}$$

Therefore, there are feasible frequency assignment and power allocation schemes such that $\alpha\mathbf{r}^{(1)} + (1-\alpha)\mathbf{r}^{(2)} \in \mathcal{C}$.

Let $p^{(1)}(f) = p^{(2)}(f)$ in the above proof. Then we have that the achievable data rate region is convex when only DSA is used. Let $D_m^{(1)} = D_m^{(2)}$ for all m in the above proof. Similarly, we have that the achievable data rate region is also convex when only APA is used. □

APPENDIX D

PROOF OF LEMMA 3.1

We only prove the case with finite channel states in this thesis. For the continuous channel state distributions, the major idea of the proof is still straightforward and very similar to that for finite channel states, but the technicality seams intricate due to measure-theoretic complications.

Proof: Let $\mathcal{J}$ represent the finite channel state set, and π_j be the stationary probability of state j, $j \in \mathcal{J}$. $T_j(t)$ denotes the subintervals of $[0,t]$ during which the channel state is j. $|T_j(t)|$ is the total length of these subintervals. Due to the ergodicity of the channel states, there exists a time t' such that for any small value $\delta > 0$,

$$\frac{|T_j(t')|}{t'} \leq \pi_j + \delta, \tag{D.1}$$

$$\text{and} \quad \liminf_{t\to\infty} \frac{\int_{\tau=0}^{t} \mathbf{r}(\tau)\, d\tau}{t} \leq \frac{\int_{\tau=0}^{t'} \mathbf{r}(\tau)\, d\tau}{t'} + \delta.$$

Thus,

$$\liminf_{t\to\infty} \frac{\int_{\tau=0}^{t} \mathbf{r}(\tau)\, d\tau}{t} \leq \sum_{j\in\mathcal{J}} \frac{|T_j(t')|}{t'} \frac{1}{|T_j(t')|} \int_{\tau\in T_j(t')} \mathbf{r}(\tau)\, d\tau + \delta.$$

According to (3.20), there exists a stationary policy $\mathcal{R}(j)$ such that

$$\frac{1}{|T_j(t')|} \int_{\tau\in T_j(t')} \mathbf{r}(\tau)\, d\tau \leq \mathcal{R}(j). \tag{D.2}$$

It follows from (D.1) and (D.2) that

$$\begin{aligned}\liminf_{t\to\infty} \frac{\int_{\tau=0}^{t} \mathbf{r}(\tau)\, d\tau}{t} &\leq \sum_{j\in\mathcal{J}} (\pi_j + \delta)\mathcal{R}(j) + \delta \\ &= \sum_{j\in\mathcal{J}} \pi_j \mathcal{R}(j) + \delta(|R(j)| + 1).\end{aligned}$$

Since $\sum_{j\in\mathcal{J}} \pi_j \mathcal{R}(j) \in \tilde{\mathcal{C}}$, let $\delta \to 0$, then

$$\liminf_{t\to\infty} \frac{\int_{\tau=0}^{t} \mathbf{r}(\tau)\, d\tau}{t} \in \tilde{\mathcal{C}}.$$

□

APPENDIX E

PROOF OF LEMMA 3.3

Proof: Let

$$\mathbf{r}^*(\mathbf{H}) = \arg \max_{\mathbf{r} \in \mathcal{C}(\mathbf{H})} \mathbf{w}^T \mathbf{r}.$$

In other words,

$$\mathbf{w}^T(\mathbf{r}_1 - \mathbf{r}^*(\mathbf{H})) \leq 0, \quad \mathbf{r}_1 \in \mathcal{C}(\mathbf{H}), \tag{E.1}$$

$$\mathbf{w}^T(\mathbf{r}_2 - \mathbf{r}^*(\mathbf{H})) \leq 0, \quad \mathbf{r}_2 \in \mathcal{C}(\mathbf{H}). \tag{E.2}$$

Let $\mathbf{r}' = \alpha \mathbf{r}_1 + (1-\alpha)\mathbf{r}_2$, where $\alpha \in (0,1)$. Then $\mathbf{r}'$ is in the convex hull of $\mathcal{C}(\mathbf{H})$. Because of (E.1) and (E.2), it follows that

$$\mathbf{w}^T(\mathbf{r}' - \mathbf{r}^*(\mathbf{H})) \leq 0, \quad \mathbf{r}' \in cov(\mathcal{C}(\mathbf{H})).$$

Taking expectation on both sides, we have

$$\mathbf{w}^T(\tilde{\mathbf{r}}' - \tilde{\mathbf{r}}^*) \leq 0, \quad \tilde{\mathbf{r}}' \in \tilde{\mathcal{C}},$$

which is equivalent to

$$\tilde{\mathbf{r}}^* = \arg \max_{\tilde{\mathbf{r}} \in \tilde{\mathcal{C}}} \mathbf{w}^T \tilde{\mathbf{r}}.$$

□

APPENDIX F

PROOF OF LEMMA 3.4

Proof:

$$\begin{aligned} Q_i[n] - \bar{Q}_i[n] &= Q_i[n] - \{(1-\rho_w)\bar{Q}_i[n-1] + \rho_w Q_i[n]\} \\ &= (1-\rho_w)\{Q_i[n] - \bar{Q}_i[n-1]\}. \end{aligned} \tag{F.1}$$

Define

$$\begin{aligned} \xi_i'[n] &= Q_i[n] - Q_i[n-1] \\ &= -\min(Q_i[n-1], r_i[n]T_s) + a_i[n]. \end{aligned} \tag{F.2}$$

Form (F.1), we have the following recurrence formula for $Q_i[n] - \bar{Q}_i[n]$,

$$Q_i[n] - \bar{Q}_i[n] = (1-\rho_w)\{Q_i[n-1] - \bar{Q}_i[n-1]\} + (1-\rho_w)\xi_i'[n]. \tag{F.3}$$

Since $\bar{Q}_i[0] = Q_i[0]$, it follows from the recursive relationship in (F.3) that

$$Q_i[n] - \bar{Q}_i[n] = \sum_{j=0}^{n-1} (1-\rho_w)^{n-j} \xi_i'[j+1].$$

We have $|Q_i[n] - \bar{Q}_i[n]| \leq \sum_{j=0}^{n-1} (1-\rho_w)^{n-j} |\xi_i'[j+1]|$, and

$$\mathbb{E}\{|Q_i[n] - \bar{Q}_i[n]|\} \leq \sum_{j=0}^{n-1} (1-\rho_w)^{n-j} \mathbb{E}\{|\xi_i'[j+1]|\}.$$

It follows from (F.2) that

$$\begin{aligned} \mathbb{E}\{|\xi_i'[j]|\} &\leq \mathbb{E}\{r_i[j]\}T_s + \mathbb{E}\{a_i[j]\} \\ &\leq (R_{\text{total}} + \lambda_i)T_s \\ &< \infty, \end{aligned}$$

where R_{total} is the maximum expected sum capacity of the system. Obviously, for $n < \infty$, $\mathbb{E}\{|Q_i[n] - \bar{Q}_i[n]|\}$ is bounded. Therefore, we need to consider the asymptotic case in which $n \to \infty$. In this case,

$$\begin{aligned}
\mathbb{E}\{|Q_i[n] - \bar{Q}_i[n]|\} &\leq (R_{\text{total}} + \lambda_i)T_s \lim_{n\to\infty} \sum_{j=0}^{n-1} (1-\rho_w)^{n-j} \\
&= \frac{1-\rho_w}{\rho_w}(R_{\text{total}} + \lambda_i)T_s \\
&< \infty.
\end{aligned}$$

□

APPENDIX G

PROOF OF THEOREM 5.1

Proof: According to the results of extreme value theory in Section 5.1, we have to show that

$$\lim_{r\to\infty}\frac{d}{dr}\left[\frac{1-F_R(r)}{f_R(r)}\right]=0,$$

if

$$\lim_{\gamma\to\infty}\frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right]=0. \tag{G.1}$$

Since

$$\frac{1-F_R(r)}{f_R(r)}=\frac{1-F_\Gamma(T^{-1}(r))}{f_\Gamma(T^{-1}(r))\left(T^{-1}\right)'(r)},$$

we have

$$\begin{aligned}
&\frac{d}{dr}\left[\frac{1-F_R(r)}{f_R(r)}\right]\\
&=-1-\frac{\left[1-F_\Gamma(T^{-1}(r))\right]\left[f'_\Gamma(T^{-1}(r))((T^{-1})'(r))^2+f_\Gamma(T^{-1}(r))\left(T^{-1}\right)''(r)\right]}{\left[f_\Gamma(T^{-1}(r))\left(T^{-1}\right)'(r)\right]^2}\\
&=\underbrace{-1-\frac{[1-F_\Gamma(T^{-1}(r))]f'_\Gamma(T^{-1}(r))}{f^2_\Gamma(T^{-1}(r))}}_{\text{Part I}}-\underbrace{\frac{[1-F_\Gamma(T^{-1}(r))]\left(T^{-1}\right)''(r)}{f_\Gamma(T^{-1}(r))\left[(T^{-1})'(r)\right]^2}}_{\text{Part II}}
\end{aligned} \tag{G.2}$$

Because $T^{-1}(r)$ is monotonically increasing with x and $T^{-1}(r)\to\infty$ as $r\to\infty$,

$$\lim_{r\to\infty}\frac{[1-F_\Gamma(T^{-1}(r))]f'_\Gamma(T^{-1}(r))}{f^2_\Gamma(T^{-1}(r))}=\lim_{\gamma\to\infty}\frac{[1-F_\Gamma(\gamma)]f'_\Gamma(\gamma)}{f^2_\Gamma(\gamma)}$$

It is easy to check that

$$\frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right]=-1-\frac{[1-F_\Gamma(\gamma)]f'_\Gamma(\gamma)}{f^2_\Gamma(\gamma)}.$$

Thus, we have

$$\lim_{r\to\infty} \text{Part I} = \lim_{r\to\infty} \frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right]. \tag{G.3}$$

Let $\tilde{T}^{-1}(r) = \dfrac{2^{\frac{x}{B}}}{\beta}$. Due to the fact that

$$\left(T^{-1}\right)''(r) = \frac{\ln 2}{B}\left(T^{-1}\right)'(r),$$

and $(\tilde{T}^{-1})'(r) = (T^{-1})'(r)$, it follows that

$$\lim_{r\to\infty} \text{Part II} = \lim_{r\to\infty} \frac{\ln 2[1-F_\Gamma(T^{-1}(r))]}{Bf_\Gamma(T^{-1}(r))\left(T^{-1}\right)'(r)} \tag{G.4}$$

$$= \lim_{r\to\infty} \frac{\ln 2[1-F_\Gamma(T^{-1}(r))]}{Bf_\Gamma(T^{-1}(r))\left(\tilde{T}^{-1}\right)'(r)}. \tag{G.5}$$

Since $\tilde{T}^{-1}(r) = T^{-1}(r) + \frac{1}{\gamma}$ and $\tilde{T}^{-1}(r) \to \infty$ as $r \to \infty$,

$$\lim_{r\to\infty} \frac{1-F_\Gamma(\tilde{T}^{-1}(r))}{f_\Gamma(\tilde{T}^{-1}(r))} = \lim_{r\to\infty} \frac{1-F_\Gamma(T^{-1}(r))}{f_\Gamma(T^{-1}(r))},$$

if (G.1) holds. Thus, we have

$$\begin{aligned}\lim_{r\to\infty} \text{Part II} &= \lim_{r\to\infty} \frac{\ln 2[1-F_\Gamma(\tilde{T}^{-1}(r))]}{Bf_\Gamma(\tilde{T}^{-1}(r))\left(\tilde{T}^{-1}\right)'(r)} \\ &= \lim_{r\to\infty} \frac{\ln 2[1-F_\Gamma(\tilde{T}^{-1}(r))]}{Bf_\Gamma(\tilde{T}^{-1}(r))\tilde{T}^{-1}(r)\frac{\ln 2}{B}} \\ &= \lim_{\gamma\to\infty} \frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)\gamma} \end{aligned} \tag{G.6}$$

Combining (G.3) and (G.6), we obtain

$$\lim_{r\to\infty} \frac{d}{dr}\left[\frac{1-F_R(r)}{f_R(r)}\right] = \lim_{\gamma\to\infty} \frac{d}{d\gamma}\left[\frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)}\right] + \lim_{\gamma\to\infty} \frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)\gamma}. \tag{G.7}$$

According to L'Hospital's rule, for a function $g(x)$ such as $g(x) \to \infty$ as $x \to \infty$, if $\lim_{x\to\infty} g'(x) = 0$, then $\lim_{x\to\infty} \dfrac{g(x)}{x} = 0$. Equation (G.1) results in

$$\lim_{\gamma\to\infty} \frac{1-F_\Gamma(\gamma)}{f_\Gamma(\gamma)\gamma} = 0,$$

Therefore, we obtain

$$\lim_{r\to\infty} \frac{d}{dr}\left[\frac{1-F_R(r)}{f_R(r)}\right] = 0.$$

Since

$$\begin{aligned} F_R^{-1}(x) &= T(F_\Gamma^{-1}(x)), \\ &= B\log_2(1+\beta F_\Gamma^{-1}(x)), \end{aligned}$$

we can obtain the normalizing constants (5.18) and (5.19) according to the results of extreme value theory in Section 5.1. □

APPENDIX H

PROOF OF EQUATION (5.23)

Proof: Let $X \geq 0$ be a random variable with distribution function $F(x)$ and $\mathbb{E}\{X\}$ is finite. The expected residual life of X is given by

$$\begin{aligned} \mathsf{R}(t) &= \mathbb{E}\{X - t | X \geq t\} \\ &= \frac{1}{1 - F(t)} \int_t^\infty 1 - F(x) dx. \end{aligned}$$

Theorem 2.1.3 and Lemma 2.7.2 in [22] show that if $F(x)$ is in the domain of the Gumbel distribution,

$$b_M = \mathsf{R}(a_M), \tag{H.1}$$

and

$$\lim_{t \to \infty} \frac{\mathsf{R}(t)}{t} = 0. \tag{H.2}$$

Since a_M monotonically increases with M, (H.1) and (H.2) directly indicates that

$$\begin{aligned} \lim_{M \to \infty} \frac{b_M}{a_M} &= \lim_{M \to \infty} \frac{\mathsf{R}(a_M)}{a_M} \\ &= \lim_{t \to \infty} \frac{\mathsf{R}(t)}{t} \\ &= 0 \end{aligned}$$

□

REFERENCES

[1] 3GPP TS 23.107, Quality of Service (QoS) concept and architecture, V6.2.0, 2004.

[2] 3GPP TS 25.308 V5.4.0, "High speed downlink packet access (HSPDA) overall description." (Release 5), Mar. 2003.

[3] Akyildiz, I., Altunbasak, Y., Fekri, F., and Sivakumar, R., "AdaptNet: An adaptive protocol suite for the next generation wireless Internet," *IEEE Commun. Magazine*, vol. 42, pp. 128–136, March 2004.

[4] Akyildiz, I., McNair, J., Carrasco, L., Puigjaner, R., and Yesha, Y., "Medium access protocols for multimedia traffic in wireless networks," *IEEE Network*, vol. 13, pp. 39–48, July-Aug. 1999.

[5] Altman, E., Basar, T., Jimenez, T., and N.Shimkin, "Competitive routing in networks with polynomial cost," in *Proc., IEEE INFOCOM*, pp. 1586 – 1593, Mar. 2000.

[6] Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., and Whiting, P., "Providing quality of service over a shared wireless link," *IEEE Commun. Magazine*, pp. 150–154, Feb. 2001.

[7] Andrews, M., "Instability of the proportional fair scheduling algorithm for HDR," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1422–1426, Sept. 2004.

[8] Andrews, M., Borst, S., Dominique, F., Jelenkovic, P., Kumaran, K., Ramakrishnan, K., and Whiting, P., "Dynamic bandwidth allocation algorithms for high-speed data wireless networks," tech. rep., Bell Labs Technical Memorandum, 2000.

[9] Bennett, J. and Zhang, H., "Hierarchical packet fair queueing algorithms," in *Proc. SIGCOMM*, pp. 143–156, Aug. 1996.

[10] Bertsekas, D. and Gallager, R., *Data Networks*. Prentice-Hall, 1987.

[11] Borst, S., "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc., IEEE INFOCOM 2003*, pp. 321–331, Mar. 2003.

[12] Cao, Y. and Li, V., "Scheduling algorithms in broadband wireless networks," *Proceedings of the IEEE*, vol. 89, pp. 76 – 87, Jan. 2001.

[13] Castillo, E., *Extreme Value Theory in Engineering*. Academic Press, 1988.

[14] Chen, C.-J. and Wang, L.-C., "A unified capacity analysis for wireless systems with joint antenna and multiuser diversity in Nakagami fading channels," in *Proc., IEEE Int. Conf. on Commun.*, June 2004.

[15] Chuang, J. and Sollenberger, N., "Beyond 3G: Wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Magazine*, pp. 78–87, July 2000.

[16] Cimini Jr., L., "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing," *IEEE Trans. Commun.*, vol. 33, pp. 665–675, July 1985.

[17] Dai, J. G., "On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Annals Applied Probab.*, vol. 5, pp. 49–77, 1995.

[18] David, H. A., *Order Statistics*. John Wiley & Sons, 1970.

[19] Decina, M. and Toniatti, T., "Bandwidth allocation and selective discarding for a variable bit rate video and bursty data calls in ATM networks," *Int'l J. Digital Analog Commun. Systems*, vol. 5, pp. 85–96, Apr.-June 1992.

[20] Eryilmaz, A., Srikant, R., and Perkins, J., "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Networking*, vol. 13, pp. 411– 424, Apr. 2005.

[21] Federgruen, A. and Groenevelt, H., "The greedy procedure for resource allocation problems: necessary and sufficient conditions for optimality," *Operations Research*, vol. 34, pp. 909–918, Nov.-Dec. 1986.

[22] Galambos, J., *The Asymptotic Theory of Extreme Order Statistics*. John Wiley & Sons, 1978.

[23] Goldsmith, A. J. and Chua, S. G., "Variable-rate variable-power MQAM for fading channel," *IEEE Trans. Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.

[24] Goldsmith, A. J. and Effros, M., "The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 47, pp. 219–240, Jan. 2001.

[25] Goodman, D. J. and Mandayam, N. B., "Power control for wireless data," *IEEE Personal Commun.*, vol. 7, pp. 48–54, Apr. 2000.

[26] Haaser, N. B. and Sullivan, J. A., *Real Analysis*. Van Nostrand Reinhold, 1971.

[27] Hoo, L., Tellado, J., and Cioffi, J., "FDMA-based multiuser transmit optimization for broadcast channels," in *Proc., IEEE Wireless Commun. Networking Conf.*, vol. 2, (Chicagp, IL), pp. 597–602, Sept. 2000.

[28] Jacobson, V., "Congestion avoidance and control," *Computer Communications Review*, vol. 18, no. 4, pp. 314–329, 1998.

[29] Jiang, Z., Ge, Y., and Li, Y., "Max-utility wireless resource management for best effort traffic," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 100–111, Jan. 2005.

[30] Kelly, F., "Charging and rate control for elastic traffic," *European Trans. On Telecommunications*, vol. 8, pp. 33–37, 1997.

[31] Kelly, F., Maulloo, A., and Tan, D., "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[32] Kivanc, D., Li, G., and Liu, H., "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1150–1158, Nov. 2003.

[33] Knopp, R. and Humblet, P., "Information capacity and power controlin single-cell multiuser communications," in *Proc., IEEE Int. Conf. on Commun.*, (Seattle, WA), June 1995.

[34] Kruse, R. L. and Ryba, A. J., *Data Structures and Program Design in C++*. Prentice Hall, 1999.

[35] Laroia, R., Uppala, S., and Li, J., "Designing a mobile broadband wireless access network," *IEEE Signal Proc. Magazine*, pp. 20–28, Sept. 2004.

[36] Li, L. and Goldsmith, A. J., "Optimal resource allocation for fading broadcast channels- part I: Ergodic capacity," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1083–1102, Mar. 2001.

[37] Li, Y. G., "Pilot-symbol-aided channel estimation for OFDM in wireless systems," *IEEE Trans. Veh. Tech.*, vol. 49, pp. 1207 –1215, July 2000.

[38] Li, Y. G., Chuang, J. C., and Sollenberger, N. R., "Transmitter diversity for OFDM system and its impact on high-rate data wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1233–1243, July 19999.

[39] Liu, P., Berry, R., and Honig, M., "Delay-sensitive packet scheduling in wireless networks," in *Proc., IEEE Wireless Commun. Networking Conf.*, (New Orleans, LA), Mar. 2003.

[40] Liu, P., Honig, M., and Jordan, S., "Forward link CDMA resource allocation based on pricing," in *Proc., IEEE Wireless Commun. Networking Conf.*, vol. 2, pp. 619 –623, Sept. 2000.

[41] Liu, X., Chong, E., and Shroff, N., "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 19, pp. pp. 2053–2064,, Qct. 2001.

[42] Liu, X., Shroff, N. B., and Chong, E. K. P., "Opportunistic scheduling: An illustration of cross-layer design," *Telecommunications Review*, vol. 16, pp. 947–959, Dec. 2004.

[43] MacKie-Mason, J. K. and Varian, H. R., "Pricing congestible network resources," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1141–1149, Sept. 1995.

[44] Maniatis, S. I., Nikolouzou, E. G., and Venieris, I. S., "QoS issues in the converged 3G wireless and wired networks," *IEEE Commun. Magazine*, pp. 44–53, Aug. 2003.

[45] McKeown, N., Mekkittikul, A., Anantharam, V., and Walrand, J., "Achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol. 47, pp. 1200–1267, Aug. 1999.

[46] Meyn, S. P. and Tweedie, R. L., *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[47] Mordecai, A., *Nonlinear Programming: Analysis and Methods*. Englewood Cliffs, N.J.: Prentice-Hall, 1976.

[48] Nanda, S., Balachandran, K., and Kumar, S., "Adaptation techniques in wireless packet data services," *IEEE Commun. Magazine*, pp. 54–64, Jan. 2000.

[49] Neely, M., Modiano, E., and Rohrs, C. E., "Power allocation and routing in multibeam satellites with time varying channels," *IEEE/ACM Trans. Networking*, vol. 11, pp. 138–152, Feb. 2003.

[50] Parekh, A. and Gallager, R., "A generalized processor sharing appproach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.

[51] Pickands III, J., "Moment convergence of sample extremes," *Annals of mathematical statistics*, vol. 39, no. 3, pp. 881–889, 1968.

[52] Qin, X. and Berry, R., "Exploiting multiuser diversity for medium access in wireless networks," in *Proc., IEEE INFOCOM*, (San Francisco, CA), pp. 1084–1094, Apr. 2003.

[53] Qiu, X. and Chawla, K., "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, pp. 884–895, June 1999.

[54] Recommendation ITU-R M.1225, "Guidelines for evaluation for of radio transmission technologies for IMT-2000," 1997.

[55] Rhee, W. and Cioffi, J. M., "Increase in capacity of multiuser OFDM system using dynamic subcarrier allocation," in *Proc., IEEE Veh. Tech. Conf.*, pp. 1085–1089, 2000.

[56] Rockafellar, R. T., *Convex Analysis*. New Jersey: Princeton University Press, 1970.

[57] Rudin, W., *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[58] Sadeghi, B., Kanodia, V., Sabharwal, A., and Knightly, E., "Opportunistic media access for multirate ad hoc networks," in *Proceedings of ACM MOBICOM 2002*, (Atlanta, GA), pp. 24–35, Sept. 2002.

[59] Saraydar, C. U., Mandayam, N. B., and Goodman, D. J., "Pricing and power control in a multicell wireless data network," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1883–1892, Oct. 2001.

[60] Schwartz, M., *Information Transmission, Modulation, and Noise*. McGraw-Hill, 1990.

[61] Shah, V., Mandayam, N. B., and Goodman, D. J., "Power control for wireless data based on utility and pricing," in *Proc., IEEE PIMRC*, pp. 1427–1432, 1998.

[62] Shakkottai, S., Rappaport, T. S., and Karlsson, P. C., "Cross-layer design for wireless networks," *IEEE Commun. Magazine*, vol. 41, pp. 74–80, Oct. 2003.

[63] Shakkottai, S. and Stolyar, A. L., "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *Analytic Methods in Applied Probability*, vol. 207, pp. 185–202, 2002.

[64] Shannon, C. E., "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[65] Shenker, S., "Fundamental design issues for the future Internet," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1176–1188, Sept. 1995.

[66] Song, G. and Li, Y. G., "Adaptive subcarrier and power allocation in OFDM based on maximizing utility," in *Proc., IEEE Veh. Tech. Conf.*, vol. 2, pp. 905 –909, Apr. 2003.

[67] Song, G. and Li, Y. G., "Cross-layer optimization for OFDM wireless network – part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 625–634, March 2005.

[68] Song, G., Li, Y. G., Cimini, L. J., and Zheng, H., "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc., IEEE Wireless Commun. Networking Conf.*, Mar 2004.

[69] Song, L. and Mandayam, N. B., "Hierarchical sir and rate control on the forward link for CDMA data users under delay and error constraints," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1871–1882, Oct. 2001.

[70] Stüber, G. L., *Principles of Mobile Communication*. Kluwer, 2 ed., 2000.

[71] Tassiulas, L. and Ephremides, A., "Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, vol. 37, pp. 1936–1949, Dec. 1992.

[72] Tassiulas, L. and Ephremides, A., "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inform. Theory*, vol. 39, pp. 466–478, March 1993.

[73] TIA/EIA IS-856, "CDMA 2000: High rate packet data air interface specification." Std., Nov. 2000.

[74] Tse, D. and Hanly, S., "Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2796–2815, Nov. 1998.

[75] Tse, D. N., "Optimal power allocation over parallel Gaussian broadcast channel," in *Proc., IEEE Int. Symp. on Inform. Theory*, (Ulm, Germany), p. 27, June 1997.

[76] Viswanath, P., Tse, D. N. C., and Laroia, R. L., "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[77] Wang, X., "An FDD wideband CDMA MAC protocol with minimum-power allocation and GPS-scheduling for wireless wide area multimedia networks," *IEEE Trans. Mobile Computing*, vol. 4, pp. 16–28, Jan.-Feb. 2005.

[78] Wittneben, A., "A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation," in *Proc., IEEE Int. Conf. on Commun.*, pp. 1630–1634, June 1993.

[79] Wong, C. Y., Cheng, R. S., Letaief, K. B., and Murch, R. D., "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.

[80] Xiao, M., Shroff, N. B., and Chong, E. K. P., "A utility-based power control scheme in wireless cellular systems," *IEEE/ACM Trans. Networking*, vol. 11, pp. 210–221, Mar. 2003.

[81] Yang, L. and Alouini, M.-S., "Performance analysis of multiuser selection diversity," in *Proc., IEEE Int. Conf. on Commun.*, June 2004.

[82] Young, R. M., "Euler's constant," *Math. Gaz.*, vol. 75, pp. 189–190, 1991.

[83] Yu, W. and Cioffi, J. M., "FDMA capacity of Gaussian multiple-access channels with ISI," *IEEE Trans. Commun.*, vol. 50, pp. 102–111, Jan. 2002.

[84] Zhou, C., Honig, M. L., and Jordan, S., "Two-cell power allocation for wireless data based on pricing," in *39th Annual Allerton Conference*, (Monticello, IL), Oct. 2001.

VITA

Guocong Song was born in October 1973 in Beijing, China. He received the B.S. and M.S. degrees in Electronic Engineering from Tsinghua University, Beijing, China, in 1997 and 2000, respectively. From July 2000 to April 2001, he was a Research Staff at the State Key Lab on Microwave and Digital Communications, Tsinghua University, China. From May 2001 to August 2005, he was a Graduate Research Assistant in the Department of Electrical and Computer Engineering at the Georgia Institute of Technology. He completed his Ph.D. in Electrical and Computer Engineering at the Georgia Institute of Technology in August 2005. His research interests are in wireless communications and networking, with a current focus on cross-layer design and optimization for wireless networks, centralized and distributed resource allocation, scheduling and MAC in multi-channel networks, and QoS provisioning.